EHDOLLINEN LOGIT-MALLI JA NAIIVI BAYES-LUOKITTELIJA - Kaksi eri tapaa diskreetin valinnan päättelyyn

Aku-Ville Lehtimäki

Helsingin yliopisto
Valtiotieteellinen tiedekunta
Tilastotiede
Pro gradu -tutkielma
helmikuu 2018

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto Fakultet/Sektion – Faculty Valtiotieteellinen tiedekunta	Laitos/Institution– Department Sosiaalitieteiden laitos	
Tekijä/Författare – Author Aku-Ville Lehtimäki		
Työn nimi / Arbetets titel – Title Ehdollinen logit-malli ja naiivi Bayes-luokittelija - Kaksi eri tapaa diskreetin valinnan päättelyyn		
Oppiaine /Läroämne – Subject Tilastotiede		
Työn laji/Arbetets art – Level Pro Gradu	Aika/Datum – Month and year Helmikuu 2018	Sivumäärä/ Sidoantal – Number of pages 31 + 7 liitettä (7 sivua)

Tiivistelmä/Referat – Abstract

Diskreetillä valinnalla tarkoitetaan tilannetta, jossa valitsijan pitää valita jokin vaihtoehto äärellisestä vaihtoehtojen joukossa. Yksilön käyttäytymisen taustalla ajatellaan yleisesti olevan taloustieteellinen, individualistinen suuntaus, jonka myötä valitsija pyrkii maksimoimaan hyötynsä. Tämän lisäksi valitsijan ajatellaan olevan rationaalinen eli toimivan tiettyjen aksiomien mukaisesti. Paradigmasta riippuen valitsijan preferenssit voivat olla satunnaiset tai deterministiset ja valitsija voi valita myös vahingossa väärin, jolloin preferenssi tai sen estimaattori on satunnaismuuttuja.

Aineisto, joka kuvaa diskreettiä valintaa, kerätään siten, että valitsijalla tai valitsijoille arvotaan joukko vaihtoehtoja, jotka koostuvat eri attribuuttien tasoista. Attribuutti on ominaisuus, esimerkiksi väri, ja sen tasoja ovat esimerkiksi punainen, vihreä ja sininen. Näin yhdellä vaihtoehdolla ei voi olla saman attribuutin kahta tasoa. Toisaalta attribuuttien määrää ei ole rajoitettu. Näiden varsinaisten vaihtoehtojen lisäksi valitsijalle on tapana esittää ei mikään -vaihtoehto, jonka valitsemalla hän pääsee pois valintatilanteesta, eikä hän esimerkiksi joudu pakotettuna valitsemaan satunnaisesti jotakin vaihtoehdoista. Jokaisesta valintatilanteesta kirjataan ylös valittavina olleet vaihtoehdot sekä tieto siitä, mikä vaihtoehto valittiin.

Perinteisesti edellä kuvattua tilannetta on estimoitu ehdollisella logit-mallilla. Se on yleistetty lineaarinen malli, eikä sen avulla eri vaihtoehtojen valintatodennäköisyyksille ole mahdollista saada analyyttisia ratkaisuja. Tämän lisäksi ei mikään -vaihtoehto tuottaa sille vaikeuksia, sillä se on oikeastaan multinomiaalisen logit-mallin luokka, ja esittämällä sen attribuuttien tasot neutraaleina tasoina lopputulemana on lineaarisen riippuvuuden ongelma. Asian ratkaisemiseksi jonkinlainen simulointi on välttämätön. Tässä pro gradu -tutkielmassa ehdollisen logit-mallin rinnalle tuodaan naiivi Bayes-luokittelija, jonka avulla on mahdollista laskea analyyttiset ratkaisut valintatodennäköisyyksille sekä ottaa mukaan ei mikään -vaihtoehto yhtenä luokkana.

Kahden aineiston avulla osoitetaan, että molemmat menetelmät ennustavat yhtä hyvin, joten tämän perusteella naiivia Bayes-luokittelijaa voi käyttää siinä missä ehdollista logit-malliakin sekä lisäksi aina silloin, kun numeerinen approksimoinnin käyttäminen ei tule kysymykseen. Tämän lisäksi todetaan, että vastaajien, jotka valitsivat ei mikään -vaihtoehdon joka kohdassa, ja täten ovat mahdollisesti vähemmän kiinnostuneita tarjotuista vaihtoehdoista, poistaminen ei tee kummastakaan mallista toista parempaa, vaikkakin osumatarkkuus molempien mallien tapauksessa parani.

Avainsanat – Nyckelord – Keywords diskreetti valinta, ehdollinen logit-malli, naiivi Bayes-luokittelija,
Säilytyspaikka – Förvaringställe – Where deposited
Muita tietoja – Övriga uppgifter – Additional information

Saatesanat

Idea tähän pro gradu -tutkielmaan syntyi omasta halustani oppia lisää diskreetin valinnan päättelystä sekä kehittää sitä edelleen ainakin nimellisesti. Taloustieteessä ja käytännöllisemmin markkinointitutkimuksessa käytettävä menetelmä on kuitenkin lopulta hyvin pienen piirin kiinnostuksen kohteena akateemisessa maailmassa. Koska kuitenkin Coxin regressio on teknisesti sama asia kuin yksinkertaisen diskreetin valinnan malli, päätin pyytää ohjaajakseni dosentti Tommi Härkästä, joka suostui ohjaamaan graduani.

Tätä kirjoittaessani minulla on jo KTK, KTM sekä VTK-tutkinnot, ja näin ollen tämä on neljäs opinnäytetyöni, johon olen saanut ohjausta. En halua tietenkään tehdä diskreettiä valintaa sen suhteen, kuka on ollut paras (tai huonoin) ohjaajistani, joten totean vain, että Tommi kuuluu ehdottomasti kahden parhaan joukkoon. Jokainen vetäköön tästä omat johtopäätöksensä. Tommi on kuitenkin antanut erinomaista palautetta pitkin matkaa, kuunnellut ideoitani ja kommentoinut niitä asiantuntevasti, mutta ennen kaikkea rakentavasti. Tämän vuoksi olen oppinut paljon tutkiessani ja kirjoittaessani.

Valtiotieteellisen tiedekunnan puolelta virallisempana ohjaajani on toiminut Kimmo Vehkalahti. Ja loppumetreillä sain myös arvokkaita kommentteja emeritusprofessori Juha Alholta. Kiitos myös heille molemmille.

Tätä pro gradu -tutkielmaa ei olisi olemassa ilman aineistoja, joiden avulla toteutin analyysit. Kiitänkin nyt teitä tai sinua, joka olet aineistot minulle luovuttanut. Tämän syvällisemmin en asiaan tässä mene, sillä aineistot ovat tietysti ehdottoman luottamuksellisia, eikä niiden lähde saa paljastua. Tiedät(te) kuka/keitä olet(te).

Ja ehkäpä tässä lopuksi voin sitten kiittää kaikkia niitä, jotka ovat minua jotenkin tässä projektissa tukeneet, joko ihan keskustelun tasolla tai vaikkapa neuvoneet Visual Basic -koodien kirjoittamisessa. Ilman teitä projekti olisi ollut luonnollisesti vieläkin mutkikkaampi.

Ja ehkäpä lopuksi on aina tapa kiittää perhettä, joka tapauksessani tarkoittaa vaimoani Pauliinaa. Onhan hän ainakin ”virkansa” puolesta joutunut ottamaan osaa tähän projektiin vähintäänkin sivustaseuraajan roolissa, eikä se varmaankaan aina ole ollut niin helppoa.

Helsingissä 26. helmikuuta 2018

Aku-Ville Lehtimäki

Kuvaluettelo

Taulukkoluettelo

1 Johdanto

Tilastotieteellisesti ihmisten mielipiteiden kysyminen sekä halujen ja tarpeiden mittaaminen, ja näihin liittyvien mallien rakentaminen voi olla hyvinkin triviaalia. Survey-tutkimuksessa on perinteisesti suosittu tunnetusti esimerkiksi Likert-asteikollisia asennekysymyksiä, ja niiden kanssa toimiminen on ainakin periaatteessa hyvin vaivatonta. Niihin liitetään kuitenkin yleensä hyvin epärealistisia oletuksia, joiden perusteella tilastollinen mallintaminen toteutetaan (ks. esim. Bishop ja Herron, 2015).

Likert-asteikollisia muuttujia on käytetty yleisesti esimerkiksi vaalikoneissa, jolloin ehdokkaita määrittyvät vastauksina vaalikoneen kysymyksiin. Kysymykset on esitetty ensin ehdokkaille, minkä jälkeen äänestäjät saavat vastata kysymyksiin omien mieltymystensä mukaisesti. Lopuksi äänestäjälle esitetään omien vastausten vastaavuus eri ehdokkaiden vastausten kanssa. Mitään objektiivista mittaa mielipiteen suunnalle ja voimakkuudelle väittämän suhteen ei kuitenkaan ole olemassa.

1.1 Diskreetti valinta

Lopullinen äänestyspäätös eli ehdokkaan valinta on esimerkki diskreetistä valintatilanteesta. Toinen esimerkki samanlaisesta on jonkin arvokkaamman tuotteen tai palvelun hankinta. Yhteistä näille tilanteille on se, että päätöksentekijä voi valita vain yhden vaihtoehdon rajallisesta määrästä vaihtoehtoja tai olla tekemättä valintaa ollenkaan.

Koska valintatilanteet vaihtelevat, päätös valinnasta muodostuu vaihtoehtojen perusteella. Vaihtoehdot koostuvat puolestaan yhdestä tai useammasta ominaisuudesta eli attribuutista. Päätöksentekijän tehdessä esimerkiksi ostopäätöstä autosta, attribuutteja ovat esimerkiksi merkki, hinta, väri, polttoaineenkulutus ja niin edelleen. Päätöksentekijällä on myös mieltymyksiä eli preferenssejä attribuuttien suhteen.

1.2 Rationaalinen päätöksentekijä

Rationaalinen päätöksentekijä kuitenkin käy analyyttisesti vaihtoehdot läpi, osaa jäsentää attribuutit omalta kannalta mieluiseen järjestykseen, ja päättää lopulta valita sen vaihtoehdon, joka vastaa hänen preferenssejään eli maksimoi hänen hyötynsä (esim. Lancaster 1966). Ollakseen rationaalista päätöksenteon on noudatettava neljää aksioomaa (Von Neumann & Morgenstern 1953): *täydellisyys* (engl. completeness), *transitiivisuus*

(engl. transitivity), *jatkuvuus* (engl. continuity) ja *riippumattomuus* (engl. independence). Aksioomat määritellään[1] tarkemmin seuraavasti

1.2.1 Aksiooma 1: Täydellisyys

Olkoon A ja B mielivaltainen vaihtoehtopari. Tällöin pätee yksi ja vain yksi seuraavista:

$$A \succ B,\ B \succ A \text{ tai } A \sim B$$

1.2.2 Aksiooma 2: Transitiivisuus

Olkoon A, B ja C mielivaltainen joukko vaihtoehtoja. Tällöin pätee:

$$\text{Jos } A \succeq B \text{ ja } B \succeq C, \text{ niin on oltava } A \succeq C.$$

1.2.3 Aksiooma 3: Jatkuvuus

Olkoon A, B ja C mielivaltainen joukko vaihtoehtoja, joille pätee:

$$A \succeq B \succeq C$$

Tällöin on olemassa suhteellinen osuus $p \in [0, 1]$ siten, että:

$$p\ A + (1 - p)\ C \sim B$$

1.2.4 Aksiooma 4: Riippumattomuus

Olkoon A ja B mielivaltainen vaihtoehtopari, jolle pätee:

$$A \succeq B$$

Tällöin on olemassa vaihtoehto C ja suhteellinen osuus $p \in (0, 1]$ siten, että:

$$p\ A + (1 - p)\ C \succ p\ B + (1 - p)\ C$$

Rationaalisuuden käsite on nähtävä laajasti. Tunteisiin tai arvoihinkin perustuva päätöksenteko on tämän määritelmän mukaisesti rationaalista, mikäli se täyttää edellä esitetyt aksioomat.

Vaihtoehdoilla on päätöksentekijäkohtainen preferenssijärjestys. Kääntäen jokaiselle vaihtoehdolle voidaan antaa jokin lukuarvo, hyöty, jonka kyseisen vaihtoehdon valitseminen päätöksentekijälle tuottaa. Vaihtoehdon hyöty puolestaan riippuu vaihtoehdon attribuuttien tasoista. Mikäli valitsijan on valittava vaihtoehdoista se, joka maksimoi hänen hyötynsä, ei tähän valintaan voida vaikuttaa muuten, kuin esittelemällä vaihtoehto, jonka hyöty on vieläkin suurempi. Muussa tapauksessa täydellisyys-aksiooma ei pitäisi paikkaansa. Transitiivisuus-aksioomasta taas seuraa, että hyödyt eivät vaihtele valintatilanteesta toiseen. Näillä perusteella vaihtoehtojen järjestyksellä tai

[1] Käytetyistä notaatioista $\succ$ ilmaisee aitoa preferenssijärjestystä, $\sim$ indifferenssiä vaihtoehtojen välillä ja $\succeq$ heikkoa preferenssijärjestystä: Päätöksentekijä valitsee A:n mieluummin kuin B:n, $A \succ B$; päätöksentekijä on indifferentti A:n ja B:n välillä, $A \sim B$; päätöksentekijä valitsee A:n mieluummin kuin B:n TAI päätöksentekijä on indifferentti A:n ja B:n välillä, $A \succeq B$.

esimerkiksi hyödyltään pienemmän vaihtoehdon lisäämisellä tai poistamisella ole vaikutusta päätöksentekijän preferensseihin. Aksioomilla 3 ja 4 ei ole yhteyttä deterministiseen, diskreettiin valintatilanteeseen. Stokastisessa tilanteessa niitä kuitenkin tarvitaan.

1.3 Satunnainen hyöty tai valintavirhe

Edellisessä kappaleessa valintatilanne oli deterministinen. Rationaalisten valintojen sarja on kuitenkin otos rajattoman suuresta valintojen joukosta. Tällöin otannassa on satunnaisuutta eli päätöksentekijän todellinen preferenssi ja valinta ovat epäyhtenäisiä tai vaihtoehtojen hyödyt ovat satunnaismuuttujia (esim. Manski 1977). Tilastotieteellisessä mielessä ensimmäinen tapa ajatella edustaa frekventististä ja jälkimmäinen (tai molemmat käsittävä) tapa bayesilaista paradigmaa.

Aksioomat 3 ja 4 sisältävät suhteellisen osuuden käsitteen, joka on frekventistisen todennäköisyyskäsityksen perusteella todennäköisyys, jolla osuuteen liittyvä vaihtoehto valitaan. Kaikille vaihtoehdoille on laskettavissa valintatodennäköisyydet päätöksentekijän preferenssien perusteella. Yksittäisen vaihtoehdon valintatodennäköisyys määritellään tällöin vaihtoehdon hyödyn ja kaikkien valintajoukon vaihtoehtojen hyötyjen summan osamääränä (esim. Luce 1977).

1.4 Diskreetin valinnan tilastollinen mallintaminen

Edellä esitetty järkeily on neutraalia sen suhteen, mistä vaihtoehdot koostuvat ja minkä perusteella päätöksentekijän preferenssin muodostuvat. Tilastollisen mallintamisen näkökulmasta näihin asioihin on kuitenkin otettava kantaa. Mikäli päätöksentekijän preferenssit otetaan annettuina ja vaihtoehdot ovat yksiattribuuttisia eli luokitteluasteikollisia, riittää periaatteessa pelkkä tapahtuneiden valintojen frekvenssitarkastelu. Varsinaiset mallit voidaan jakaa kolmeen eri kategoriaan, sillä perusteella onko selittäjinä päätöksentekijän ominaisuudet, vaihtoehtojen ominaisuudet vai molemmat (esim. Duncan & Hoffman, 1988): Multinomiaalinen logit-malli soveltuu yksiattribuuttisten vaihtoehtojen mallintamiseen päätöksentekijäkohtaisten taustamuuttujien avulla. Ehdollinen logit-malli soveltuu puolestaan tilanteeseen, jossa yksi- tai useampiattribuuttisten vaihtoehtojen valintaa on tarve mallintaa vaihtoehtojen ominaisuuksilla. Sekamallin selittävinä muuttujina voi olla puolestaan sekä vaihtoehto- että päätöksentekijäkohtaisia taustamuuttujia.

1.5 Tutkimuksen tarkoitus

Edellisessä kappaleessa esitetyt mallit ovat kaikki yleistettyjä lineaarisia malleja. Tämän vuoksi numeeriset menetelmät ovat välttämättömiä valinnan ennustamiseksi (LÄHDE) On kuitenkin tilanteita ja ympäristöjä, joissa numeeriset menetelmät ovat mahdottomia tai ainakin työläitä toteuttaa. Työssä tarkastellaan ja arvioidaan mahdollisuutta mallintaa vaihtoehtojen ominaisuuksiin perustuvaa diskreettiä valintaa naiivilla Bayes-luokittelijalla verrattuna ehdolliseen logit-malliin. Vaikka naiivin Bayes-luokittelija käsitteleekin vaihtoehtoja luokkina ja tilannetekijöitä eli päätöksentekijän ominaisuuksia selittäjinä, on vaihtoehtojen ominaisuudet mahdollista esittää tilannetekijöinä keskiarvoistamalla. Tämän lisäksi tapaukset, joissa päätöksentekijä ei halua tehdä valintaa, ovat suoraviivaisempia ottaa huomioon naiivilla Bayes-luokittelijalla kuin ehdollisella logit-mallilla.

1.6 Työn rakenne

Tämä pro gradu –tutkielma on jaettu seitsemään lukuun. Tätä johdantokappaletta seuraa diskreetin valinnan koeasetelman kuvaus. Koeasetelman monimutkaisuus on varmasti yksi syy, miksi menetelmää käytetään melko vähän verrattuna perinteisempiin survey-menetelmiin. Koeasetelman kuvauksen jälkeen seuraa kappale, jossa esitellään tarkemmin naiivi Bayes-luokittelija sekä ehdollinen logit-malli. Neljännessä luvussa puolestaan esitellään estimointimenetelmät, jotka tulevat kyseeseen kummankin menetelmän kohdalla. Viidennessä luvussa esitellään aineistot, joihin perustuen menetelmiä testataan kuudennessa luvussa. Seitsemännessä luvussa keskustellaan tutkimusta reflektoiden ja kriittisiäkin näkökulmia esittäen.

2 Koeasetelma

Varmasti yksi syy siihen, miksi johdantokappaleessa esitetty Likert-asteikollinen mittaaminen on suosittua, on sen helppous. Diskreetin valinnan tapauksessa itse koeasetelmakin on monimutkaisempi.

2.1 Valintatilanne

Kokeeseen osallistuu *D* kappaletta päätöksentekijöitä (engl. decision maker):

$$d = 1, \dots, D$$

Jokainen päätöksentekijä käy läpi *m* kappaletta valintatilanteita *r*, joiden teoreettinen enimmäismäärä on R (engl. risk set):

$$r = 1, \dots, R \geq m$$

Kullekin päätöksentekijälle *d* arvotaan siis satunnaisesti jokin vakiomäärä *m* valintatilanteita kokeen aikana *R*:n valintatilanteen joukosta, siten että arvottu valintatilanne palautetaan mahdollisten valintatilanteiden joukkoon. Valintatilanteiden kokonaismäärä kokeen aikana on *D*m = M*. Jokainen valintatilanne ajatellaan lisäksi toisista valintatilanteista riippumattomaksi.

Koska tässä työssä ajatellaan, että kaikilla vastaajilla on samanlaiset preferenssit, on tutkimusongelman kannalta yhdentekevää, onko yksi vastaaja vastannut kaikkiin valintatilanteisiin eli 1**m = M* vai onko jokainen vastaaja vastannut yhteen valintatilanteeseen eli *D**1 = M*. Näiden ääripäiden väliltä myös mitkä tahansa sellaiset *D*:n ja *m*:n arvot, joiden tulo vastaa aineistonmukaista valintatilanteiden kokonaismäärää *M*, tulevat kysymykseen.

2.2 Valintatilanteen vaihtoehdot

Jokaiseen valintatilanteeseen *r* kuuluu vakiomäärä eli *A* kappaletta vaihtoehtoja (engl. alternative). Tämän vuoksi vaihtoehtojen kokonaismäärä kokeen aikana on M*A=n ja yksittäistä vaihtoehtoa merkitään *i*:

$$i = 1, \dots, n$$

Jokainen valintatilanne voidaan määritellä yksilöllisesti sen sisältämillä vaihtoehdoilla tai niitä vastaavilla asetelmavektoreilla X_i.

Vaihtoehdot koostuvat puolestaan attribuuteista *c*, joita on jokaisella vaihtoehdolla *C* kappaletta. Jokaisessa valintatilanteessa on siis esillä yhteensä *C*A=W* kappaletta yksittäisiä attribuutteja *w*:

$$w = 1, \dots, W$$

Attribuutteja voi olla mielivaltainen määrä, mutta kokeen jokaisella vaihtoehdolla valintatilanteen sisällä sekä valintatilanteiden välillä on sama määrä attribuutteja. Ja vaihtoehdon sisällä jokaisella attribuutilla voi olla kerrallaan vain yksi taso. On hyvä huomata, että edellä esitettyjen, varsinaisten vaihtoehtojen lisäksi vaihtoehtojen joukkoon kuuluu ei mikään -vaihtoehto. Sen valitsemalla päätöksentekijä ilmaisee, että ei valitsisi mitään annetuista vaihtoehdoista.

2.2.1 Ei mikään -vaihtoehto

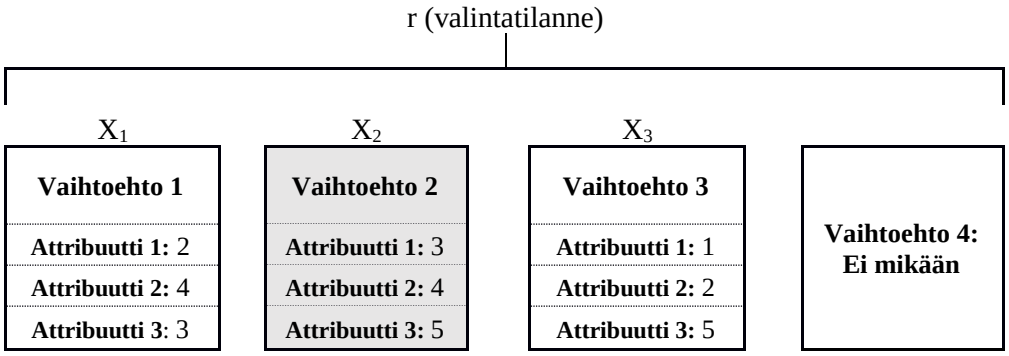
Johdannossa esitetyn viitekehyksen perusteella päätöksentekijä on rationaalinen, ja hänen on mahdollista asettaa vaihtoehdot haluamaansa preferenssijärjestykseen. Ei mikään -vaihtoehdon tarjoaminen liittyy kuitenkin yleisesti oletuksiin esimerkiksi siitä, ettei päätöksentekijä osaa käsitellä kaikkea valintatilanteessa olevaa tietoa tai on epävarma siitä, mitä haluaa. (esim. Dhar 1997). Nämä olettamat ovat toki arkijärjen mukaisia, mutta ne eivät sovi rationaalisen päätöksentekijän määritelmään. Sen sijaan rationaalisella päätöksentekijällä saa olla preferenssi nykytilan säilyttämisen puolesta, mutta on koeasetelman määrittelijän päätettävissä, onko tällainen vaihtoehto tarjolla. Todellisuudessa joissain tilanteissa päätöksentekijä voi valita nykytilan säilyttävän, neutraalin vaihtoehdon, kun taas toisissa valintatilanteissa tätä mahdollisuutta ei ole. Hyvä esimerkki neutraalin vaihtoehdon luontevuudesta on vaikkapa ostotilanne, jossa ei mikään -vaihtoehdon hinta olisi 0 € ja muilta ominaisuuksiltaan se olisi tyhjä. Käytännössä kokeeseen on voinut myös valikoitua mukaan vastaajia, joita ei sitten lopulta kiinnostakaan keskittyä kokeen tekemiseen, ja heille annetaan joustava mahdollisuus sensuroida itsensä pois.

Ei mikään -vaihtoehto on mukana tässä työssä, sillä se on mukana alkuperäisissä aineistoissa, ja on luonteva siinä kontekstissa, jossa kokeet alun perin suoritettiin. Toinen syy neutraalin vaihtoehdon sisällyttämiseen on se, että sitä käsitellään hyvin eri tavalla naiivilla Bayes -luokittelijan ja ehdollisen logit-mallin tapauksessa tuoden eroavaisuutta mallien välille.

2.2.2 Esimerkki valintatilanteesta

Kuvassa 1 on kuvattu kuvitteellinen valintatilanne, jossa on kolme varsinaista vaihtoehtoa sekä ei mikään -vaihtoehto, ja valitsija on valinnut vaihtoehdon 2.

Kuva 1: Diskreetti valintatilanne



Yleisesti jokainen valintatilanne arvottuine vaihtoehtoineen sekä valittu vaihtoehto kirjataan ylös. Tuntemattomaksi jää, mikä olisi ollut valitsemattomien vaihtoehtojen valintajärjestys. Edellä esitetystä tilanteesta olisi siis kirjattu ylös, että vaihtoehdon *i* attribuuttien 1, 2 ja 3 tasot ovat 2, 4 ja 3, vaihtoehdon *a* attribuuttien tasot ovat 3, 4 ja 5 jne. Tämän lisäksi kirjataan muistiin, että kyseisessä valintatilanteessa valittiin vaihtoehto *a*.

Attribuuttien tasoilla ei ole olemassa mitään loogista järjestystä, sillä päätöksentekijän kokeman hyödyn ja attribuutin muutoksen välillä ei ole olemassa mitään loogisesti johdettavaa (monotonista) funktiota. Tämän vuoksi jokaisen attribuutin taso on oma luokkansa. Teknisesti vaihtoehdolla voisi olla useampiakin saman attribuutin tasoja, mutta attribuuttikohtaiset vektorit on määritelty niin, että niissä vain yksi alkio saa arvon 1 ja loput arvon 0.

Kun kuvaa 1 katsotaan nyt dikotomisten selittäjien näkökulmasta, ja olkoon attribuutilla 1 kolme tasoa, attribuutilla 2 neljä tasoa ja attribuutilla 3 viisi tasoa, niin nyt vaihtoehdolle 1 voidaan muodostaa seuraavanlainen asetelmavektori, jossa selvyyden vuoksi attribuuttien rajat eli attribuuttikohtaiset vektorit on erotettu puolipisteillä:

$$X_1 = (0,1,0;\ 0,0,0,1;\ 0,0,1,0,0)$$

Helposti voisi ajatella, että ei mikään -vaihtoehdolle voidaan määritellä samankaltainen asetelmavektori kuin muillekin vaihtoehdoille. Tämä ei kuitenkaan käy, sillä mikäli jokaiselle attribuutille määrittelisi ei mikään -tason, niin tällöin attribuutit riippuisivat

tämän tason osalta toisistaan, ja lopputuloksena olisi lineaarisen riippuvuuden tilanne. Tämä olisi taas loogisesti vastoin sitä oletusta, että attribuuttien tasot ovat toisistaan riippumattomia.

3 Mallien kuvaus

Tässä luvussa esitellään ensin ehdollinen logit-malli, joka koostuu kahdesta osasta: systemaattisesta ja satunnaisesta. Tämän jälkeen siirrytään naiiviin Bayes-luokittelijaan. Molempien mallien osalta oletetaan, että vaihtoehto on aina hyötyjensä summa.

3.1 Ehdollinen logit-malli

Olkoon Y_r diskreetti valinta joukosta r, jossa on A kappaletta vaihtoehtoja. Olkoon U_i hyöty, jonka päätöksentekijä saa i:nnen vaihtoehdon valinnasta. Nyt hyödyn U_i voi jakaa esimerkiksi lineaarisesti systemaattiseen osaan η_i sekä satunnaiseen osaan ε_i:

$$U_i = \eta_i + \varepsilon_i$$

Johdanto-kappaleessa esitellyn rationaalisen päätöksentekijän määritelmän mukaisesti rationaalinen päätöksentekijä valitsee keskimäärin i:nnen vaihtoehdon, mikäli sillä on joukon r suurin hyöty. Koska valinta on satunnainen, määritellään valintatodennäköisyys, että päätöksentekijä valitsee vaihtoehdon i seuraavasti:

$$p_i = P(Y_r = i) = P(\max\{U_i, \dots, U_A\} = U_i)$$

Virhetermit ovat toisistaan riippumattomia ja noudattavat puolestaan tyypin 1 ääriarvojakaumaa eli:

$$f(\varepsilon_i) = f(\varepsilon) = \exp\{-\varepsilon - \exp\{-\varepsilon\}\}$$

Näiden perusteella voidaan johtaa (ks. esim. Maddala 1983, s. 60 – 61):

$$p_i = \frac{exp(\eta_i)}{\sum_{j=1}^{A} exp(\eta_j)}$$

Huomataan, että todennäköisyys on sama kuin johdantokappaleessa esitetty Lucen (1977) esittämä.

3.2 Naiivi Bayes-luokittelija

Naiivi Bayes-luokittelija (engl. naive Bayes classifier, NBC) on ehdollisen todennäköisyyden malli, joka perustuu Bayesin kaavaan. Naiivius puolestaan viittaa oletukseen, että selittävät muuttujat eivät riipu toisistaan. Toisaalta diskreetin valinnan koeasetelmassa koeasetelman toteuttaja voi määrittää, etteivät vaihtoehtojen attribuuttien tasot todella riipu toisistaan, ei oletus toki ole ollenkaan naiivi. Naiivin Bayes-luokittelijan etuna voidaan pitää sitä, että luokkien todennäköisyyksille on olemassa

analyyttinen ratkaisu, eikä iteratiivisia menetelmiä tarvitse käyttää (esim. Russell & Norvig 2009, s. 499, 505 ja 808).

Itse malli määritellään seuraavasti: Olkoon L_k luokka k ja vektori $\mathbf{z} = (z_1, \ldots, z_W)$ tilanteen ominaisuuksia eli selittäjiä kuvaava vektori. Ehdolliseen logit-malliin vertautuvassa diskreetin valinnan tilanteessa selittäjät ovat tietenkin vaihtojen ominaisuuksia toki kuitenkin siten, että järjestyksellä on väliä.

Nyt todennäköisyys, että näiden tiettyjen ominaisuuksien vallitessa tilanne luokitellaan luokkaan L_k, voidaan merkitä:

$$P(L_k|z_1, \ldots, z_W) = P(L_k|\boldsymbol{z})$$

Nyt Bayesin kaavana nojalla voidaan merkitä:

$$P(L_k|\boldsymbol{x}) = \frac{P(L_k)P(\boldsymbol{z}|L_k)}{P(\boldsymbol{z})}$$

Oikeanpuolen osoittajan oikeanpuolimmainen termi voidaan kirjoittaa ketjutussäännön perusteella:

$$P(L_k)P(\boldsymbol{z}|L_k) = P(z_1|z_2, \ldots, z_W, L_k)P(z_2, \ldots, z_W, L_k)$$

$$= P(z_1|z_2, \ldots, z_W, L_k)P(z_2|z_3, \ldots, z_W, L_k) \ldots P(z_{W-1}|z_W, L_k)P(z_W|L_k)P(L_k)$$

Koska selittäjät ovat toisistaan riippumattomia, voidaan kirjoittaa:

$$P(z_w|x_{w+1}, \ldots, x_W, L_k) = P(x_w|L_k)$$

Nyt alkuperäinen malli voidaan tulona seuraavasti:

$$P(L_k|\boldsymbol{x}) = \frac{P(L_k)P(\boldsymbol{x}|L_k)}{P(\boldsymbol{x})} = \frac{P(L_k)\prod_{w=1}^{W}P(x_w|L_k)}{P(\boldsymbol{x})}$$

Nyt määritelty malli ei kuitenkaan ole yhtenäinen johdantoluvun viitekehyksen kanssa, sillä sen mukaan vaihtoehtojen järjestyksellä ei ole väliä. Tämä ristiriita otetaan huomioon luvussa 4.3 vaihtoehtojen valintatodennäköisyyksiä estimoitaessa.

4 Estimointimenetelmät

Edellisessä luvussa esiteltiin mallit. Seuraavaksi esitellään molempiin malleihin liittyvät estimointimenetelmät. Lisäksi kerrotaan, miten vastaamatta jättämistä eli ei mikään -vaihtoehdon valintaa tulisi estimoida.

4.1 Ehdollinen logit-malli ja suurimman uskottavuuden estimointi

Malleja esiteltäessä esiteltiin ehdollinen logit-malli, jossa kyse on jonkin systematiikan vaikutuksesta valintatodennäköisyyteen:

$$p_i = \frac{exp(\eta_i)}{\sum_{j=1}^{A} exp(\eta_j)}$$

Kaavan oikeanpuolen nimittäjässä summattavat termit j = 1, …, A kuuluvat samaan valintajoukkoon kuin osoittajan i. Kun oletetaan, että valinnat ovat toisistaan riippumattomia ja valintatilanteita on m kappaletta, joissa jokaisessa on n vaihtoehtoa, voidaan mallin uskottavuusfunktio määritellä seuraavasti:

$$\prod_{i=1}^{n} \left(\frac{exp(\eta_i)}{\sum_{j \in r} exp\left(\eta_j\right)} \right)^{\delta_i}$$

Missä δ_i saa arvon 1, kun vaihtoehto i valittiin, ja muuten arvon 0.

Mikäli selittäjät ovat dikotomisia, kuten ne olivat kappaleen 2.2.2. esimerkissä, voidaan systemaattinen osa kirjoittaa asetelmavektorin ja kerroinvektorin pistetulona:

$$\eta_i = X_i \beta$$

Jokaista attribuutin tasoa vastaa oma kerroin. Uskottavuusfunktio voidaan puolestaan kirjoittaa muodossa:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{exp(X_i \beta)}{\sum_{j \in r} exp(X_j \beta)} \right)^{\delta_i}$$

Missä ainoastaan kerroinvektori β ajatellaan muuttujana ja kaikki muut termit vakioina, kun koe on suoritettu ja aineisto kerätty. Suurimman uskottavuuden estimoinnin tavoitteena on sellainen kerroinvektori β, jolla uskottavuusfunktio maksimoituu. Uskottavuusfunktio maksimoituu samassa kohdassa kuin logaritminen uskottavuusfunktio eli log-uskottavuusfunktio, mutta näistä jälkimmäinen on laskennallisesti yksinkertaisempi käsitellä. Log-uskottavuusfunktio on muotoa:

$$logL(\beta) = l(\beta) = log \prod_{i=1}^{n} \left(\frac{exp(X_i\beta)}{\sum_{j \in r} exp(X_j\beta)} \right)^{\delta_i} = \sum_{i=1}^{n} \delta_i \left(X_i\beta - log \sum_{j \in r} exp(X_j\beta) \right)$$

Mikäli funktio on yleisesti kaikkialla derivoituva, sen ääriarvokohta on derivaatan nollakohdassa. Koska kyseessä on nyt useamman muuttujan funktio eli parametreja on useita, olemme kiinnostuneita parametrien arvoista, joilla gradientti on nollavektori. Gradientti koostuu osittaisderivaatoista kaikkien β-vektorin alkioiden suhteen, eli:

$$\nabla l(\boldsymbol{\beta}) = (\frac{\delta l(\boldsymbol{\beta})}{\delta \beta_0}, \frac{\delta l(\boldsymbol{\beta})}{\delta \beta_1} \dots)^T = \mathbf{0}$$

Mitään tunnettua, yleistä ja analyyttistä ratkaisua $\boldsymbol{\beta}$-vektorin alkioiden arvoille gradientin nollakohdassa ei ole olemassa. Yhtälö voidaan kuitenkin ratkaista esimerkiksi Newtonin-Raphsonin menetelmällä (esim. Agresti 2013, s. 143). Olkoon $\nabla l(\boldsymbol{\beta})$ gradienttivektori kuten yllä ja olkoon $\boldsymbol{H}$ vastaava Hessen matriisi, joka koostuu alkioista h_{ab}:

$$h_{ab} = \frac{\delta^2 l(\boldsymbol{\beta})}{\delta \beta_a \delta \beta_b}$$

Missä a ja b viittaavat riviin a ja sarakkeeseen b. Olkoon $\nabla l(\boldsymbol{\beta})^{(t)}$ ja $\boldsymbol{H}^{(t)}$ lasketut gradientti ja Hessen matriisi, kun kyseessä on t:nnes arvio $\widehat{\boldsymbol{\beta}}$:n arvoksi eli $\boldsymbol{\beta}^{(t)}$. Iteratiivisen prosessin kierros t (t = 0, 1, 2, 3, …) approksimoi $l(\boldsymbol{\beta})$:n $\boldsymbol{\beta}^{(t)}$:n läheisyydessä, kun toisen kertaluvun Taylorin polynomi on:

$$l(\boldsymbol{\beta}) \approx l(\boldsymbol{\beta}^{(t)}) + \nabla l(\boldsymbol{\beta})^{(t)^T} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) + 0{,}5 (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^T \boldsymbol{H}^{(t)} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})$$

Seuraava arvio saadaan ratkaisemalla $\boldsymbol{\beta}$:n suhteen:

$$\nabla l(\boldsymbol{\beta}) = \nabla l(\boldsymbol{\beta})^{(t)} + \boldsymbol{H}^{(t)} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) = \mathbf{0}$$

Olettaen, että $\boldsymbol{H}^{(t)}$on ei-singulaarinen, itse arvio on puolestaan:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - (\boldsymbol{H}^{(t)})^{-1} \nabla l(\boldsymbol{\beta})^{(t)}$$

Iteraatioita jatketaan, kunnes kahden peräkkäisen arvion ero on tarpeeksi pieni.

4.2 Ehdollisen logit-mallin kerroinestimaattorin kovarianssimatriisi

Aiemmin todettiin, että hyödyt ovat satunnaisia, joten estimaattori noudattaa useampiulotteista jakaumaa. Estimaattorin kovarianssimatriisi saadaan maksimoimalla osittaisen log-uskottavuusfunktion Hessen matriisi eli (ks. esim. Sy & Taylor, 2001):

$$\frac{\delta^2}{\delta\beta^2} l(\beta) = -\sum_{i=1}^{n} \delta_i \left\{ \frac{\sum_{j\in r} X_j X_j' exp(X_j\beta)}{\sum_{j\in r} exp(X_j\beta)} - \frac{\left(\sum_{j\in r} X_j exp(X_j\beta)\right)\left(\sum_{j\in r} X_j exp(X_j\beta)\right)'}{\left(\sum_{j\in r} exp(X_j\beta)\right)^2} \right\}$$

4.3 Naiivi Bayes-luokittelija

Naiivi Bayes-luokittelija on Bayesin kaavan sovellus, ja todennäköisyyksien laskeminen liittyy käänteistodennäköisyyksiin.

Diskreetin valinnan tilanteessa sitä voidaan soveltaa niin, että valinnan i todennäköisyys on ehdollinen valintajoukon r_{jk} suhteen:

$$p_{i|qk} = P\left(vaihtoehto\ i \mid joukko\ r_{qk}\right) = \frac{P\left(joukko\ r_{qk} \mid vaihtoehto\ i\right) P(vaihtoehto\ i)}{P\left(joukko\ r_{qk}\right)}$$

Uskottavuusfunktio on Bayesin kaavan tapauksessa käänteistodennäköisyyksien tulo:

$$\frac{P\left(joukko\ r_{qk} \mid vaihtoehto\ i\right)}{P\left(joukko\ r_{qk}\right)} = \frac{\prod_{q=1}^{A!} exp\left(X_i \alpha_{i|qk}\right)}{\sum_{j\in R_{qk}} \prod_{q=1}^{A!} exp\left(X_i \alpha_{i|qk}\right)} = \frac{exp\left(\sum_{q=1}^{A!} X_i \alpha_{i|qk}\right)}{\sum_{j\in R_{qk}} exp\left(\sum_{q=1}^{A!} X_i \alpha_{i|qk}\right)}$$

missä q viittaa tiettyyn vaihtoehtojen järjestykseen A! järjestysten kokonaismäärään. Nämä kaikki käänteistodennäköisyydet ovat ehdollisia todennäköisyyksiä, joissa ehtona on kyseisen vaihtoehdon valinta. Aineistosta valitaan kaikki ne tapaukset, kun kyseinen vaihtoehto valittiin. Tämän jälkeen aineistosta katsotaan niiden tilanteen ominaisuuksien todennäköisyyksiä eli ilmenemisfrekvenssejä, jotka ovat kiinnostuksen kohteena. Ehdolliseen logit-malliin verrattaessa näitä ovat tietenkin eri vaihtoehtojen attribuutit. Uskottavuusfunktio valinnalle *i* ehdolla järjestetyt vaihtoehdot muodostetaan siis kertomalla kaikkien näiden järjestettyjen vaihtoehtojen attribuuttien todennäköisyydet keskenään tarkasteltaessa aineistossa, jossa on vain mukana tapaukset, kun *i* valittiin. Jokainen näistä todennäköisyyksistä eli kerrottavista termeistä on peräisin binomiprosessista.

Ehdollisen logit-mallin tapauksessa oletetaan aina, että järjestyksellä ei ole väliä, ja ainoat vaikuttimet vaihtoehtojen valintojen suhteen ovat todellakin vain ja ainoastaan

vaihtoehtoihin liittyvät hyödyt. NBC:n tapauksessa ei tarvitse olla tätä mieltä, ja binomiprosessien parametrien erisuuruutta voi testata esimerkiksi asiaankuuluvalla tilastollisella testillä. Jos kuitenkin tekee sen päätöksen, ettei järjestyksellä ole väliä, niin kahdelle vaihtoehtojoukolle, jossa on samat vaihtoehdot mutta eri järjestyksessä, pitäisi vastaaville vaihtoehdoille tulla keskimäärin samat todennäköisyydet eli:

$$p_{i|qk} = p_{a|ql},$$

missä *i* viittaa vaihtoehtoon ja *a* viittaa samoista attribuutttien tasoista koostettuun vaihtoehtoon. Vaihtoehdot *i* ja a sijaitsevat tyypin *q* valintajoukossa (r_q) ja niiden kanssa kilpailevat vaihtoehdot ovat myös identtisiä, mutta nyt valintajoukkojen vaihtoehtojen järjestys onkin eri (*k* ja *l*). Havaitut poikkeamat eri valintajoukkojen tosiaan vastaavien vaihtoehtojen valintatodennäköisyyksien eli piste-estimaattien välillä voivat siis johtua ainoastaan satunnaisvaihtelusta. Lopulliset, järjestyksestä riippumattomat todennäköisyyksien estimaatit saadaan laskettua poimimalla jokaisesta tyypin *q* valintajoukosta toisiaan attribuuttien tasoiltaan vastaavat alkiot, ja laskemalla näiden keskiarvot eli:

$$\frac{p_{i|qk} + p_{a|ql} + \cdots}{A!}$$

missä A! viittaa järjestysten lukumäärään. Tämä toistetaan kaikille eri vaihtoehdoille, jolloin saadaan laskettua kullekin vaihtoehdolle valintatodennäköisyys.

4.4 Ei mikään -vaihtoehto

Naiivin Bayes-luokittelijan tapauksessa ei mikään -vaihtoehto on yksinkertaisesti yksi luokista, johon selittävien muuttujien perusteella tilanteen voi mahdollisesti luokitella.

Ehdollinen logit-malli perustuu teknisesti suoraan siihen, etteivät selittävät muuttujat eli attribuutit korreloi keskenään ja jokaista vaihtoehtoa määrittävät samat attribuutit. Mikäli ei mikään -vaihtoehto määritellään vakioiduilla attribuuttien tasoilla, attribuuttien korreloimattomuus ei ole voimassa.

Ongelman kiertämiseksi on käyty akateemista keskustelua, mutta siihen ei ole olemassa yhtä oikeaa ratkaisua (esim. Haaijer, Kamakura & Wedel, 2001). Käytän tässä pro gradu -tutkielmassa kaksivaiheista mallia ei mikään -vaihtoehdon hyödyn estimoimiseen ehdollisen logit-mallin tapauksessa.

Ensimmäisessä vaiheessa estimoidaan hyödyt ehdolla, jokin varsinaisista vaihtoehdoista valittiin. Käytännössä tämä tarkoittaa ehdollisen logit-mallin parametrien estimointia aineistolla, josta on poistettu valintatilanteet, joissa ei mikään -vaihtoehtoehto tuli valituksi.

Toinen vaihe puolestaan alkaa sillä, että lasketaan kaikkien niiden tapausten lukumäärä, kun ei mikään -vaihtoehto valittiin. Näiden valintatilanteiden lukumäärää merkitään *n(ei mikään)*, kun taas kaikkien valintatilanteiden lukumäärää merkitään *n(kaikki)*. Olemme kiinnostuneita Ei mikään -vaihtoehdon hyödystä:

$$U_{ei\ mikään}$$

joka on tuntematon reaaliluku. Lucen (1977) valinta-aksiooman mukaisesti, voimme muodostaa seuraavan yhtälön:

$$\frac{n(ei\ mikään\ valittiin)}{n(kaikki\ tilanteet)} = \frac{exp\{U(ei\ mikään)\}}{exp\{U(ei\ mikään)\} + exp\{U(keskimääräinen\ maksimi)\}}$$

$$\Leftrightarrow U(ei\ mikään) = \log\frac{n(ei\ mikään)}{n(kaikki)} - U(keskimääräinen\ maksimi) + \log(1 - \frac{n(ei\ mikään)}{n(kaikki)})$$

Ylläolevassa merkinnät ovat yhteneviä kappaleen 3.1 merkintöjen kanssa eli simulointikokeessa on kaksi vaihtoehtoa: Ei mikään ja valintatilanteen maksimi. *U(keskimääräinen maksimi)* saadaan, kun käydään *n* kappaletta valintatilanteita läpi ja lasketaan jokaisesta tilanteesta maksimihyöty eli:

$$\frac{1}{n}\sum_{i=1}^{n} U_i(valintatilanteen\ maksimi)$$

Kun oletetaan, että hyödyt ovat ainakin keskimääräisesti kiinteitä, voidaan keskimääräinen maksimihyöty simuloida arpomalla suuri määrä valintatilanteita (esimerkiksi 1 000 000 kappaletta) eli:

$$\frac{1}{1\ 000\ 000}\sum_{i=1}^{1\ 000\ 000} U_i(valintatilanteen\ maksimi)$$

Jokaisesta valintatilanteesta valitaan aina vaihtoehto, jolla on suurin hyöty. Näiden eri kierrosten maksimihyödyistä lasketaan lopuksi keskiarvo.

5 Aineistojen kuvaus

Tutkielmaa varten olen saanut kaksi erilaista aineistoa, joista molemmat ovat peräisin oikeista kyselyistä, jotka ovat osana kaupallisten toimijoiden tiedonhankintaa. Kaupallisten toimijoiden kaiken toiminnan perimmäisenä tarkoituksena on tuloksen maksimointi. Jokainen päätös, joka tällaisen organisaation tai henkilön vuoksi tehdään, on siis odotusarvoltaan suotuisin rahallisessa mielessä. Tästä tutkimuksesta ei ole kerta kaikkiaan mitään hyötyä tälle toimijalle edes teoriassa (pl. goodwill-arvo eli se positiivinen maine, joka minun suuntaani kasvaa), mutta periaatteessa riskinä on, että tutkimuskohde selviää muille toimijoille. Riskiä lisää vielä entisestään se, ettei tämänkaltaisia tutkimuksia tehdä edes yleismaailmallisesti kovin paljoa verrattuna esimerkiksi perinteisiin kyselytutkimuksiin.

Näiden lähtökohtien vallitessa olen saanut aineistot käyttööni sillä ehdolla, että muokkaan attribuuttien ja niiden tasojen nimet täysin erilaisiksi, siten että alkuperäisiä aineistoja on mahdoton tai ainakin erittäin vaikea tunnistaa. Tämä naamiointi ei kuitenkaan vaikuta millään käsityksessäni olevalla tavalla analyysien oikeellisuuteen, minkä vuoksi on perusteltua väittää tämän julkisuustason olevan riittävä tilastotieteen pro gradu -tutkielmaan.

Aineistot olivat alun perin tekstimuodossa siten, että muuttujien arvot oli erotettu toisistaan pilkulla. Jokainen vaihtoehto oli omana havaintoyksikkönä omalla rivillään. Muina muuttujina olivat se, että mistä valintatilanteesta on kyse, sekä se, että mikä vaihtoehto valittiin. Ei mikään -vaihtoehtoa varten oli oma muuttujansa, joka sai arvon 0 kaikkien niiden valintatilanteiden havaintoyksikköjen kohdalla, jolloin jokin varsinaisista vaihtoehdoista valittiin. Muutoin tämä muuttuja sai arvon 1. Aineistossa oli myös mukana vastaajakohtainen tunniste sekä muuttuja, joka kertoi, että mikä oli minkäkin vaihtoehdon paikka kyseisessä valintatilanteessa. Taulukossa 1 on esitetty kuvitteellisella aineistolla edellä mainittu muoto.

Vastaaja	Valintatilanne	Attribuutti_1	…	Attribuutti_n	Valittiin	Ei mikään valittiin	Paikka
1	1	3	…	4	1	0	2
1	1	2	…	3	0	0	4
…							
5	68	2	…	2	0	1	1
5	69	1	…	3	0	1	2
…jne.							

Taulukko 1: Alkuperäisten aineistojen muoto

Ensimmäisessä aineistossa on kyse kissanruokakonsepteista. Attribuutteja on kolme kappaletta: maku, koostumus ja hinta. Maku-attribuutti jakautuu viiteen eri tasoon: lohi, tonnikala, kana, häränliha ja grillattu liha. Koostumuksia on kolmenlaisia: hyytelö, kuiva sekä stick. Mahdollisia hintoja on viisi erilaista: 1) 8,99 € 2) 5,30 € 3) 4,29 € 4) 3,90 € ja 5) 2,50 €.

Toisessa aineistossa kyse on puolestaan älypuhelinkonsepteista, jotka ovat teknisesti samoja tai melko samanlaisia. Kolme attribuuttia, joiden tasojen mieluisuudesta on vastaajilta kysytty, ovat: väri, näytön koko sekä hinta. Väriattribuutilla on viisi tasoa: punainen, vihreä, sininen, musta ja valkoinen. Näytön koon tasoina ovat 4 tuumaa, 5 tuumaa ja 6 tuumaa. Hinta-attribuutti voi olla: 1) 299 €, 2) 199€ ja 3) 99 €. Yhteenveto aineistoista on esitetty taulukossa 2.

Aineisto	Valintatilanteiden lukumäärä	Vastaajien lukumäärä	Valintatilanteita vastaajaa kohden	Vaihtoehtojen lukumäärä	Attribuuttien tasojen lukumäärä
1) Kissanruoka	1356	113	12	4 + *ei mikään*	5 + 3 + 5
2) Älypuhelin	2250	250	9	3 + *ei mikään*	5 + 3 + 3

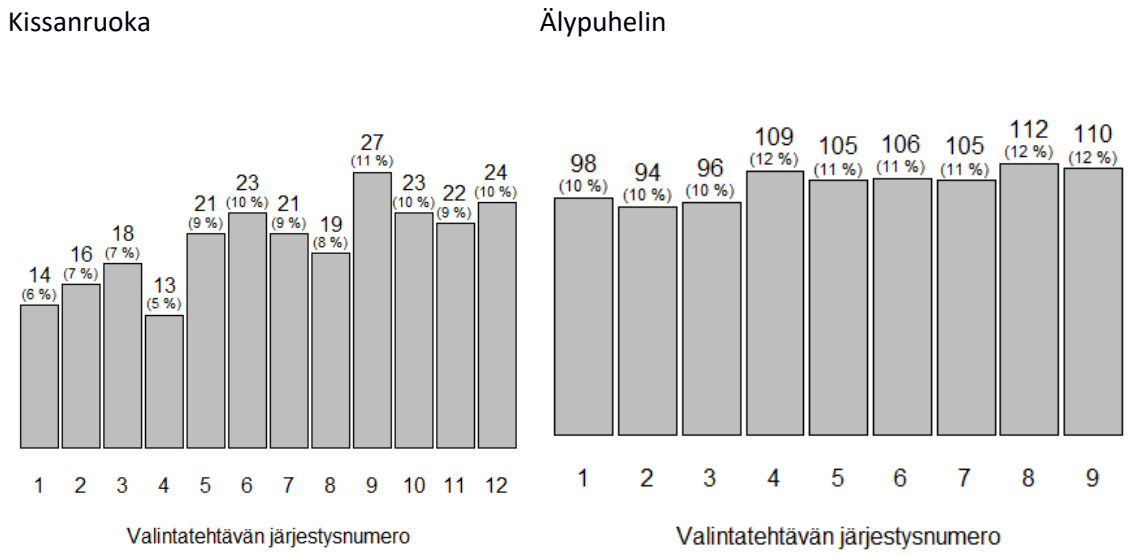
Taulukko 2: Aineistojen koeasetelmien kuvaukset

Kuten taulukosta 2 nähdään, kissanruokaa varten vastaajia oli vähemmän kuin älypuhelinta varten, samoin valintatilanteita. Toisaalta vastaajakohtaisia valintatilanteita on enemmän kissanruokakonsepteista kyseltäessä.

Kaikki kuvatut attribuutit ovat luokitteluasteikollisia, sillä niiden preferenssijärjestystä ei ole mielekästä johtaa mistään muusta loogisesta järjestyksestä, johon ne voidaan asettaa. Esimerkiksi hinta-attribuutin suhteen tämä tarkoittaa sitä, että vaikka korkeampi hinta aiheuttanee varmaankin haittaa jokaiselle vastaajalle enemmän kuin alhaisempi hinta menetetyn rahan muodossa, korkeampi hinta saattaa toimia mahdollisesti myös korkeamman laadun indikaattorina.

Molemmista aineistoista on tiedossa vastaajakohtaisesti valintatilanteiden järjestys. Järjestys on satunnainen. Valintatilanteen indeksinumeron perusteella on mahdollista vertailla kasvaako alttius valita ei mikään -vaihtoehto valintatilanteiden määrän kasvaessa tai muuttuuko tämä alttius yleisestikään valintatilanteiden järjestysnumeroiden välillä. Tilanne on esitetty graafisesti kuviossa 2.

Kuva 2: Ei mikään -vaihtoehdon valinnan jakautuminen valintatilanteiden välillä



Nollahypoteesin pitäessä paikkaansa luokkien (eli valintatilanteiden) ja luokkakohtaisten havaintojen lukumääristä laskettu testisuure noudattaa khii-toiseen-jakaumaa vapausastein ”valintatehtävien lukumäärä” – 1 (Pearson 1900).

Khii-toiseen-yhteensopivuustestien perusteella havaitut p-arvot ovat 0,557 ja 0,922 kissanruoka- ja älypuhelinaineistoille tässä järjestyksessä (ks. laskemiseen käytetty R-koodi tuloksineen liitteessä 1). Voidaan siis pitää kiinni oletuksesta, että vastaajien ei voida katsoa muuttavan alttiutta valita ei mikään -vaihtoehtoa valintatilanteiden järjestysnumeron muuttuessa.

6 Mallien testaus

Mallien testauksella ja vertailulla selvitetään yleisesti mallien kykyä ennustaa. Kokonaisuudessaan valintatilanteet:

$$i = 1, \dots, n$$

muodostavat opetusaineiston (engl. training set) **r**. Tässä mielessä jokainen havaintoyksikkö koostuu kahdesta osasta:

$$r = (t_r, y_r)$$

missä t_r on selittäjien vektori ja y_r puolestaan vastemuuttuja. Perustuen opetusaineistoon voidaan muodostaan ennustesääntö:

$$\eta(t, \boldsymbol{r})$$

Tarkoituksena on tähän sääntöön nojaten ennustaa tuntematon havainto y_0. Vastemuuttuja olkoon dikotominen ja olkoon $Q[y_r, \eta_r]$ ennusteen oikeellisuus siten, että:

$$Q[y_r, \eta_r] = \begin{cases} 0, kun\ \eta_r = y_r \\ 1, kun\ \eta_r \neq y_r \end{cases}$$

missä $\eta_r = \eta(t_r, \boldsymbol{r})$ viittaa ennusteeseen, joka on vastemuuttujan tavoin tietenkin myös dikotominen. Nyt ennustesäännön todellinen virhesuhde (Err; engl. true error rate) voidaan määritellä todennäköisyytenä ennustaa tuntematon havainto (t_0, y_0) väärin. Toisin sanoen voidaan puhua odotusarvosta:

$$E\{Q[Y_0, \eta(t_r, \boldsymbol{r})]\}$$

Kiinnostuksen kohteena on nyt estimoida todellinen virhesuhde opetusaineiston perusteella. Selvästikin virheellisten ennusteiden suhteellinen osuus eli ilmeinen virhesuhde (engl. apparent error rate) eli:

$$\frac{1}{n}\sum_{i=1}^{n} Q[y_i, \eta(t_i, \boldsymbol{r})]$$

mittaa tätä. Tarkemmin kyse on siis siitä, kuinka paljon suhteellisesti ennustevirheitä ennustesäännön perustella ilmenee aineistossa, jonka perusteella ennustesääntö on muodostettu, eli opetusaineistossa. Tämä estimaatti on kuitenkin usein pienempi kuin todellinen, ennustesääntöön liittyvä todellinen virhesuhde. Tämä siksi, että ennustesääntö sekä muodostetaan että arvioidaan saman aineiston perusteella. (esim. Efron 1993).

6.1 Ristiinvalidointi

Efronin (1993) mukaan ongelma, jossa ilmeinen virhesuhde on keskimäärin pienempi kuin todellinen virhesuhde, voidaan kiertää ristiinvalidoinnilla. Yleisesti ideana on poistaa aineistosta vuorotellen jokainen havainto ennustesääntöä muodostettaessa, ja käyttää tätä vajaalla aineistolla muodostettua ennustesääntöä poistetun havaintoyksikön vastetta ennustettaessa. Muodollisesti voidaan siis sanoa, että olkoon $\boldsymbol{r}_{(i)}$ opetusaineisto, josta havaintoyksikkö $\boldsymbol{r}_i$ on poistettu ja olkoon $\eta(t, \boldsymbol{r}_{(i)})$ vastaava ennustesääntö. Nyt ristiinvalidoitu virhesuhde on:

$$\frac{1}{n}\sum_{i=1}^{n} Q[y_r, \eta(t_i, \boldsymbol{r}_{(i)})].$$

Edellä esitettyä jätä-yksi-pois (engl. leave-one-out, LOO) -menetelmää on kritisoitu mm. korkeahkosta varianssista (esim. McLachlan, Do, Ambroise, 2004, s. 214). McLachlan ym. (2004, s. 214) esittelevätkin q-lohkoisen (engl. q-fold) ristiinvalidoinnin, jossa opetusaineisto jaetaan lohkoihin, joita on q kappaletta ja joiden koko on m havaintoyksikköä, eli:

$$n = q * h, h \geq 1$$

Olkoon opetusaineisto, josta on poistettu g:s lohko, nyt:

$$\boldsymbol{r}_{(g)} = (\boldsymbol{r}'_1, \dots, \boldsymbol{r}'_{(g-1)h}, \boldsymbol{x}'_{gh+1}, \dots, \boldsymbol{x}'_n)'$$

Nyt q-lohkoinen, ristiinvalidoitu virhesuhde on:

$$\frac{1}{n}\sum_{i=1}^{h}\sum_{j=1}^{q} Q[y_{(r)}, \eta(t_{(j-1)h+i}, \boldsymbol{r}_{(j)})]$$

McLahlanin ym. (2004, s. 214) mukaan liian pienillä q:n arvoilla osaopetusaineistot ovat liian pieniä suhteessa koko opetusaineistoon, ja arvot esimerkiksi q:n arvo 10 on hyvä kompromissi jätä-yksi-pois-menetelmän ja liian suurien osaopetusaineistojen välillä.

Koska kissanruoka-aineistossa valintatilanteiden lukumäärä ei ole jaollinen luvulla 10, on viimeinen ryhmä tämän aineiston kohdalla hieman pienempi havaintoyksiköillä mitattuna.

Ristinvalidoinnissa jokainen havaintoyksikkö tarvitsee siis oman tunnisteensa eli kokonaisluvun väliltä 1 – 10 ehdolla, että jokaista tunnistetta on (suunnilleen) yhtä monta

kappaletta ja että jokainen tunniste on muutoin osoitettu satunnaisesti jokaiselle havaintoyksikölle.

Aineisto laitettiin satunnaiseen järjestykseen luomalla satunnaisia arvoja, järjestämällä aineisto generoitujen arvojen mukaan luomalla sarake, jossa sekvenssi 1-10 toistui n/10 kertaa, ja tämän jälkeen palauttamalla aineisto alkuperäiseen järjestykseen. Koska kissanruoka-aineistossa valintatilanteita oli 1356 kappaletta, eikä luku ole jaollinen luvulla 10, toistettiin mainittua sekvenssiä n+4 / 10 = 1360 / 10 kertaa. Viimeinen ryhmä jäi siis neljää havaintoa pienemmäksi kuin muut ryhmät.

Kissanruoka-aineiston suhteen osumatarkkuus on logit-mallilla 478 / 1356 eli noin 35,3 % ja älypuhelinaineistolla 815 / 2250 eli noin 36,2 %. Kun sellaiset vastaajat poistetaan, jotka ovat valinneet ei mikään -vaihtoehdon, nousevat osumatarkkuudet suhteeseen 478 / 1296 eli noin 36,9 %:iin ja suhteeseen 815 / 1692 eli noin 48,2 %:iin tässä järjestyksessä.

Edellä mainittiin siitä ongelmasta, jonka sinänsä hyödyllinen ei mikään -vaihtoehto tuo tullessaan: Vastaaja voi valita sen, jos ei vain jaksa keskittyä tehtävään kunnolla. Ongelmaan vastattiin siten, että mallit testattiin myös aineistoilla, joista oli poistettu ne vastaajat, jotka olivat valinneet jokaisessa valintatilanteessa ei mikään -vaihtoehdon. Aineistoa kuvailtaessa suljettiin pois myös se testaamalla, että alttius valita ei mikään -vaihtoehto ei muutu vastaajakohtaisesti valintatilanteiden välillä.

Molemmissa tapauksissa jokaisen havainnon kohdalle saatiin todennäköisyydet eri vaihtoehdoille. Ennusteeksi valittiin se vaihtoehto, jolle oli ennustettu suurin todennäköisyys.

6.2 Ehdolliset logit-mallit

Kahta aineistoa varten estimoidaan kaksi ehdollista logit-mallia. Molemmissa malleissa ensin estimoitiin varsinaisten vaihtoehtojen attribuuttien tasojen hyödyt, minkä jälkeen siirryttiin estimoimaan ei mikään -vaihtoehdon hyötyjä. Estimointi oli siis kaksivaiheinen.

6.2.1 Ensimmäinen vaihe: Parametrien estimointi

Ensin estimoitiin parametrit ehdolla ”valitsija valitsi jonkin varsinaisista vaihtoehdoista”. Vaihe toteutettiin käyttäen IBM SPSS 25 -ohjelman lisäosaa Complex Samples, joka mahdollistaa kertoimien estimoimisen Newtonin(-Raphsonin) menetelmällä. Aluksi aineisto jaettiin kymmeneen eri ryhmään aiemmin mainitun ristiinvalidointiryhmittelyn

mukaisesti: Mikäli ryhmän tunniste oli k, niin siihen ryhmään kuului kaikki ne havainnot, joiden indikaattori oli jokin muu kuin k. Koska SPSS-ohjelman versiossa 25 ei ole proseduuria ehdollisen logit-mallin parametrien estimointiin, käytettiin Coxin regressiota. Coxin regressio seurattavine tutkimussubjekteineen ja ehdollinen logit-malli vaihtoehtoineen ovat laskennallisesti sama asia, kun jokainen valintatilanne ajatellaan omana seurantajaksonaan ja seuranta lopetetaan ensimmäisen tapahtuman ilmentymiseen eli valinnan tapahtumiseen, ja loput tutkimussubjektit tai vaihtoehdot sensuroidaan. Tämän lisäksi ensimmäisen tapauksen ilmenemiseen kulunut aika vakioidaan tilanteiden välillä (esim. Le & Lindgren, 1988). Viimeinen ehto tarvitsee SPSS 25 -ohjelmassa oman aikamuuttujansa, joka on oikeastaan vakio, sillä se saa saman arvon jokaisen havaintoyksikön osalta. Estimaatit sekä niihin liittyvät keskivirheet on esitetty liitteessä 2.

6.2.2 Toinen vaihe: Ei mikään -vaihtoehto

Käyttäen IBM SPSS 25 -ohjelmaa laskettiin ei mikään -vaihtoehtojen suhteellinen ilmenemismäärä jokaisessa 10 ryhmästä. Koska tiedossa oli edellisestä vaiheesta sekä odotusarvon että varianssin estimaatit, oli mahdollista simuloida valintatilanteita. Tätä varten rakennettiin Microsoft Excel 2016 -ohjelmalla laskentataulukko, johon arvottiin RAND()-funktiolla erilaisia mahdollisia valintatilanteita siten, että jokaisessa näkyi jokaisen vaihtoehdon attribuuttien tasojen hyödyt. Nämä hyödyt laskettiin yhteen, ja niistä valittiin suurin. Taulukkolaskentaympäristössä jokaisen arvonnan tuloksen sijoittaminen omaan soluunsa olisi kömpelöä, piti työkirjaa laajentaa Visual Basic -komentosarjalla, jossa valintatilanteiden arvontaa toistettiin 1 000 000 kertaa silmukalla. Jokaisella kierroksella jokaisen ristiinvalidointiryhmän hyödyllisimmän vaihtoehdon hyöty tallennettiin muistiin, ja lisättiin edellisten kertojen vastaavan summaan. Lopulta, kun komentosarja oli silmukan osalta valmis, jaettiin tämä summa luvulla 1 000 000. Simulointiin käytetty Visual Basic -komentosarja on esitetty liitteessä 3, ja ositekohtaiset ei mikään -vaihtoehdon hyödyt on esitetty liitteessä 2.

6.3 Naiivit Bayes-luokittelijat

Koska Naiivi Bayes-luokittelija on luokittelija, ei alkuperäinen havaintomatriisin muoto ole perusteltu. Alkuperäiset aineistot avattiin Microsoft Excel 2016 -ohjelmaan, ja muokattiin muotoon, jossa vastemuuttujana oli vaihtoehdon järjestys alkuperäisessä valintatilanteessa, ja muina muuttujina oli vaihtoehtojen attribuutit, ja nämä muuttujat saivat arvoikseen vastaavat attribuuttien tasot.

Vastaaja	Vaihtoehdon paikka valintatilanteessa	Attribuutti_1	…	Attribuutti_n
1	2	3	…	4
1	4	2	…	3
…				
5	1	2	…	2
5	2	1	…	3
…jne.				

Taulukko 3: Aineiston muoto naiivin Bayes-luokittelijan tapauksessa

Naiivi Bayes-luokittelija laskee syötteiden eli annettujen attribuuttien tasojen perusteella todennäköisyydet luokille eli järjestetyille vaihtoehdoille. Tätä varten piti rakentaa erillinen Excel-työkirja, jossa jokaisen attribuutin tason todennäköisyys ehdolla ”tietty vaihtoehto valittiin” poimittiin aineistosta erikseen. Apuna tässä oli Microsoft Excel -ohjelman tietokantafunktiot – lähinnä DCOUNT(). Eri järjestykset laskettiin eri välilehtiin siten, että jokaista eri järjestystä vastasi yksi välilehti. Näiden eri välilehtien valitsimia kontrolloitiin yhdestä välilehdestä eli päävälilehdestä siten, että siellä annetut attribuuttien tasot menivät järjestyksien mukaisille paikoilleen eri järjestyksiä edustavissa välilehdissä. Tällä tavoin pystyttiin laskemaan jokaiselle järjestykselle omat todennäköisyydet, joista sitten laskettiin keskiarvot päävälilehteen oikeille paikoilleen.

Jokaiselle havainnolle määritettiin sama ristiinvalidointiryhmä kuin ehdollisten logit-mallienkin tapauksessa. Ja tässä mielessä myös estimointi noudatti samaa kaavaa eli jokaista havaintoa ennustettiin niiden havaintojen perusteella, jotka eivät kuuluneet havainnon kanssa samaa ristiinvalidointiryhmään. Toimenpiteen koodi on esitetty liitteessä 3.

Naiivin Bayes-luokittelijan osumatarkkuus on nyt kissanruoka-aineistolla 474 / 1356 ≈ 35,0 % ja älypuhelinaineistolla 815 / 2250 ≈ 36,2 %. Kun poistetaan vastaajat, jotka valitsivat joka kerralla ei mikään -vaihtoehdon osumatarkkuudet ovat nyt 473 / 1296 ≈ 36,5 % ja 815 / 1692 ≈ 48,2 % samassa järjestyksessä. Tässä kohdin on huomionarvoista mainita, että älypuhelinaineistolla mallien välillä ei ole ollenkaan eroa ennustamisen suhteen.

6.4 Mallien vertailu

Mallin osumatarkkuuden on oltava vähintään vaihtoehtojen lukumäärän käänteisluku, jotta voidaan sanoa, että malli havaitsee edes jotain systematiikkaa aineistosta. Kissanruoka-aineiston tapauksessa tämä luku on 1/5 = 0,2 ja älypuhelimen tapauksessa 1/4 = 0,25. Mikäli osumatarkkuus on tätä pienempi, malli on systemaattisesti harhainen.

Mikäli osumatarkkuus on yhtä suuri kuin edellä mainittu tunnusluku, ei malli havaitse mitään systematiikkaa aineistosta. On hyvä huomioida, että aineisto voi olla itsessään niin satunnainen, ettei edes perusteltu ja järkevästi rakennettu malli pysty havaitsemaan siinä mitään systematiikkaa. Näin on esimerkiksi, mikäli vastaajat eivät ole jaksaneet keskittyä kunnolla valintaan, vaan valinneet jonkin vaihtoehdoista aina satunnaisesti. Tämän vuoksi mallien vertailussa onkin käytetty kahta eri aineistoa.

6.4.1 Ehdollinen logit-malli

Jokaisen ristiinvalidointiryhmän attribuuttien tasojen kertoimet sekä ei mikään -vaihtoehdon kerroin taulukoitiin, ja ne kopioitiin Microsoft Excel -työkirjaan. Tämän jälkeen valintatilanteen attribuuttien tasojen perusteella laskettiin todennäköisyydet ja ennuste koskien jokaista havaintoa. Samalla laskettiin, että millä vaihtoehdolla oli suurin todennäköisyys. Mikäli tämä vaihtoehto oli sama kuin se, joka oli oikeastikin valittu, kirjattiin havainnon kohdalle osuma.

6.4.2 Naiivi Bayes-luokittelija

Koska naiiviin Bayes-luokittelijaan tulee syöttää aina kaikkien vaihtoehtojen attribuuttien tasot, jonka jälkeen todennäköisyydet lasketaan, ja havaintoja oli melko paljon, piti työtä automatisoida. Tätä tarkoitusta varten kirjoitettiin Visual Basic -komentosarja, joka syötti havainnon vaihtoehtojen attribuuttien taso päävälilehteen, ja kopioi sieltä lasketut todennäköisyydet ennusteiksi. Komentosarja siirtyi tämän jälkeen seuraavaan havaintoon, ja jatkoi tätä, kunnes koko aineisto oli käyty läpi. Komentosarja on esitetty liitteessä 4. Samalla tavalla kuin ehdollisen logit-mallin tapauksessa myös naiivin Bayes-luokittelijan kohdalla verrattiin vaihtoehtoihin liittyvien todennäköisyyksien maksimiarvoa valitsijan tekemään valintaan, ja mikäli nämä olivat yhtenevät, havainto kirjattiin osumaksi.

6.4.3 Tilastolliset testit

Kokonaisuutena kiinnostuksen kohteena on se, että milloin malli ennustaa oikein (tai väärin). Selittävinä muuttujina on malli (NBC ja logit), aineisto (kissanruoka ja älypuhelin) ja tilanteet, joissa kaikki vastaajat ovat mukana ja toisaalta tilanteet, joissa jokaisessa tilanteessa ei mikään -vaihtoehdon valinneet vastaajat on poistettu. Kiinnostus on kuitenkin mallikohtaisessa eikä aineistokohtaisessa suoriutumisessa. Ei mikään -vaihtoehdon suhteen kyseessä on taas menetelmäkohtainen ominaisuus tai haaste.

Tämän perusteella voidaan muodostaa seuraavanlaiset hypoteesit: 1) Kumpi malli ennustaa paremmin? 2) Paraneeko mallin ennustavuus, mikäli kaikki sellaiset vastaajat poistetaan, jotka ovat valinneet ei mikään -vaihtoehdon?

6.4.4 Kumpi malli ennustaa paremmin?

Ensimmäistä tutkimusongelmaa voidaan lähestyä McNemarin (1947) testillä. Yleisesti se soveltuu tilanteeseen, jossa ollaan kiinnostuneita siitä, kuinka samalla tavoin kaksi eri dikotomista prosessia realisoituvat samojen havaintoyksikköjen kohdalla. Tässä tilanteessa kyse on siis siitä, että kuinka samalla tavoin eri menetelmät ennustavat, ja nollahypoteesi on tietysti se, että menetelmät ennustavat samalla tavalla oikeiden ja väärien ennusteiden mielessä. Mahdolliset ennusteparit sekä niiden esiintyminen aineistossa on ristiintaulukoitu taulukossa 4.

Kissanruoka

		Ehdollinen logit -malli		
		OE	VE	Yht
NBC	OE	466	8	474
	VE	12	870	882
	Yht	478	878	1356

Älypuhelin

		Ehdollinen logit -malli		
		OE	VE	Yht
NBC	OE	815	0	815
	VE	0	1435	1435
	Yht	815	1435	2250

Kaikki

		Ehdollinen logit -malli		
		OE	VE	Yht
NBC	OE	1281	8	1289
	VE	12	2305	2317
	Yht	1293	2313	3606

Taulukko 4: Oikeiden ja väärien ennustusten yhtenevyys mallien välillä; OE = oikea ennuste, VE = väärä ennuste

McNemarin testin oletus on se, että testisuure on asymptoottisesti khii-toiseen-jakautunut yhdellä vapausasteella, kun nollahypoteesi pitää paikkansa. Testisuure perustuu ideaan, että ristiriitatilanteita on kahdenlaisia 1) NBC ennusti oikein ja logit väärin 2) NBC ennusti väärin ja logit oikein. Näiden lukumäärien erotuksen neliö skaalataan ristiriitatilanteiden kokonaismäärällä eli:

$$\frac{(8-12)^2}{12+8} = \frac{16}{20} = \frac{4}{5} = 0{,}8$$

Ennen kuin testisuureen arvosta vetää johtopäätöksiä, on hyvä huomata, että tapauksia, joissa mallit ennustavat toinen oikein ja toinen virheellisesti eli ristikkäistapauksia, on absoluuttisesti melko vähän. Tässä mielessä khii-toiseen jakauma ei ole välttämättä kovin hyvä referenssijakauma nollahypoteesin tapauksessa. Edellä laskettua testisuuretta vastaava p-arvo on noin 0,371 ja tarkan p-arvon voi laskea käyttämällä referenssijakaumana binomijakaumaa. Nollahypoteesina tarkassa tapauksessa on se, että molempia ristikkäistapauksia ilmenee yhtä paljon, eli kaikista ristikkäistapauksista molempia on 50 %. Yleisesti tarkan p-arvon voi laskea kaavalla:

$$\sum_{i=0}^{a}\binom{a+b}{i}0{,}5^{i}0{,}5^{a+b-i}=\sum_{i=b}^{a+b}\binom{a+b}{i}0{,}5^{i}0{,}5^{a+b-i}=\frac{p}{2}$$

Kaavassa a ja b viittaavat ristikkäistapausten lukumääriin siten, että:

$$a \leq b$$

Näiden summa on luonnollisesti kaikkien ristikkäistapausten lukumäärä. Havaittu p-arvo eli p pitää jakaa kahdella, sillä kyseessä on kaksisuuntainen testi. Binomikertoimen ominaisuuksiin kuuluu, että yllä kuvatulla tavalla summattuina ne ovat todellakin yhtä suuret. Johtuen jakauman symmetrisyydestä summan binomikertoimen jälkeinen osa ei riipu indeksistä, joten tarkan p-arvon kaava voidaan kirjoittaa muodossa:

$$0{,}5^{a+b-1}\sum_{i=0}^{a}\binom{a+b}{i}=p$$

Kun saatujen ristikkäistapausten lukumäärä asetetaan kaavaan, saadaan tarkaksi p-arvoksi noin 0,503. Kummalla tahansa tavalla voidaan vetää se johtopäätös, ettei mallien välillä ole eroa, kun riskitasoksi on valittu mikä tahansa yleisesti hyväksytty taso. Edellä esitetyt tulokset on laskettu R-ohjelman funktioilla mcnemar.test ja binom.test, ja syntaksit listauksineen on esitetty liitteessä 6.

Taulukossa 5 on esitetty ristiinvalidointiryhmien vaihteluvälit eri mallien ja aineistojen välillä.

Kissanruoka

Malli	**Alaraja**	**Yläraja**
Ehdollinen logit-malli	0,28	0,45
Naiivi Bayes-luokittelija	0,28	0,43

Älypuhelin

Malli	**Alaraja**	**Yläraja**
Ehdollinen logit-malli	0,33	0,4
Naiivi Bayes-luokittelija	0,33	0,4

Taulukko 5: Ristiinvalidointiryhmien vaihteluvälit eri mallien ja aineistojen välillä

Kuten taulukosta 5 näkyy ei mallien välillä ole käytännössä eroa. Tälläkin perusteella voidaan siis ajatella, että mallit ovat tasaväkisiä.

6.4.5 Ei mikään -vaihtoehdon joka kerran valinneiden poistaminen

Kun aineistosta poistetaan kaikki ne vastaajat, jotka valitsivat ei mikään -vaihtoehdon jokaisessa valintatilanteessaan tilanne ei käytännössä muutu mihinkään, eli ristikkäisennusteet pysyvät miltei ennallaan. Uuden ristiintaulukot on esitetty taulukossa 6.

Kissanruoka

		Ehdollinen logit -malli		
		OE	VE	Yht
NBC	OE	466	7	473
	VE	12	811	823
	Yht	478	818	1296

Älypuhelin

		Ehdollinen logit -malli		
		OE	VE	Yht
NBC	OE	815	0	815
	VE	0	877	877
	Yht	815	877	1692

Kaikki

		Ehdollinen logit -malli		
		OE	VE	Yht
NBC	OE	1281	7	1288
	VE	12	1688	1700
	Yht	1293	1695	1988

Taulukko 6: Oikeiden ja väärien ennustusten yhtenevyys mallien välillä, kun joka kerran ei mikään -vaihtoehdon valinneet on poistettu; OE = oikea ennuste, VE = väärä ennuste

Taulukosta 6 nähdään, että kissanruoka-aineistossa ehdollinen logit-malli on kirinyt yhdessä tapauksessa naiivia Bayes-luokittelijaa kiinni, mutta tällaisen muutoksen voi tämän kokoisessa aineistossa toki arvioida jo ennalta mahtuvan satunnaisvaihtelun piiriin. Tämä siitäkin huolimatta, että McNemarin testisuure muuttuu melko radikaalisti:

$$\frac{(7-12)^2}{12+7} = \frac{25}{19} = 1\frac{6}{19} \approx 1{,}32$$

Khii-toiseen jakaumaa käytettäessä referenssijakaumana McNemarin testisuureeseen liittyvä havaittu p-arvo on nyt 0,25 ja binomijakaumaan perustuva tarkka p-arvo on puolestaan 0,359. Nämä tulokset vahvistavat ylempänä mainittua ensivaikutelmaa siitä, että mallien keskinäinen paremmuus ei muutu. Taulukossa 7 vielä on esitetty ristiinvalidointiryhmien vaihteluvälit eri mallien ja aineistojen välillä.

Kissanruoka

Malli	**Alaraja**	**Yläraja**
Ehdollinen logit-malli	0,30	0,47
Naiivi Bayes-luokittelija	0,30	0,45

Älypuhelin

Malli	**Alaraja**	**Yläraja**
Ehdollinen logit-malli	0,42	0,52
Naiivi Bayes-luokittelija	0,42	0,52

Taulukko 7: Ristiinvalidointiryhmien vaihteluvälit eri mallien ja aineistojen välillä, kun joka kerran ei mikään -vaihtoehdon valinneet on poistettu

Mallit ovat tasaväkisiä edelleen. Lopputulemaan ei näytä vaikuttavan, vaikka jokaisessa valintatilanteessa ei mikään -vaihtoehdon vastanneet vastaajat on poistettu.

7 Johtopäätökset ja yhteenveto

Tässä pro gradu -tutkielmassa verrattiin ehdollista logit-mallia ja naiivia Bayes-luokittelijaa diskreetin valinnan ennustamisen mielessä. Viitekehyksenä taustalla oli rationaalisen toimijan olettama. Koska naiivi Bayes-luokittelija on luokittelumalli, täytyy sillä lasketuista luokittelutodennäköisyyksistä keskiarvoistaa valintatodennäköisyydet vaihtoehdoille, jotka viitekehyksen mukaisesti ovat attribuutteihinsa liittyvien hyötyjen lineaarikombinaatioita.

Mallien välillä ei ollut eroa ennustamisen mielessä. Älypuhelinaineistolla ennusteet olivat jopa identtisiä, eikä kissanruoka-aineistollakaan osumatarkkuudessa ollut mallien välillä tilastollisesti merkitsevää eroa. Tulos saattaakin olla analoginen sen kanssa, että generatiiviset luokittelijat (ml. naiivi Bayes-luokittelija) ja diskriminatiiviset luokittelijat eli logistisen regression mallit toimivat samalla periaatteella, kun mallin selittäjät ovat dikotomisia (esim. Ng & Jordan 2002). Tämän huomioiden eroa olisi voinut kuvitella vääristävän ei mikään -vaihtoehto. joka vaati ehdollisen logit-mallin laajentamisen simuloinnilla.

Jatkoa ajatellen mielenkiintoinen tutkimuskohde voisikin olla jatkuvien selittäjien käyttäminen, sillä ne pystyvät ottamaan huomioon selittävien muuttujien jakaumat naiivin Bayes -luokittelijan tapauksessa. Viitekehys ei kuitenkaan aseta yksilöiden preferensseille mitään monotonista yhteyttä attribuuttien muutoksen suhteen, joten tässä kontekstissa jatkuvien selittäjien käyttäminen on yleisestikin haasteellista. Esimerkiksi korkeaa hintaa on ikävä maksaa, mutta samalla se voi olla viesti korkeasta laadusta. Ainakin aluksi mahdolliset lokaalit ääriarvokohdat lienevät perusteltua selvittää dikotomisilla selittäjillä.

Tutkielmassa oli oletuksena, että jokaisella päätöksentekijällä on samanlaiset preferenssit. Tähän ei johdantokappaleessa esitelty viitekehys tietenkään pakota, mutta oletus oli tehty siitä syystä, että vastaajakohtaisia havaintoja oli melko vähän attribuuttien tasojen lukumäärään verrattuna. Asia on siinä mielessä tärkeä, että jo ihan arkikokemuksenkin perusteella voidaan todeta, etteivät kaikki ihmiset pidä samoista asioista. Jos esimerkiksi meillä on kaksi henkilöä, joista toinen pitää jääteestä (+2 C°) ja toinen kuumasta teestä (+80 C°), niin mikäli esimerkiksi lineaarisesti interpoloidaan heidän kokonaishyötynsä maksimoituvan heidän molempien saadessa teetä, joka on +41 C°:ista, niin lopputulos voikin oikeasti olla kaukana kokonaishyödyn maksimoitumisesta.

Haaste on toki tiedostettu kirjallisuudessa, ja siihen on ehdotettu ja kehitetty ohjaamattomaan oppimiseen perustuvia ratkaisuja, kuten piilevien luokkien analyysi (engl. Latent Class Analysis, LCA; esim. Greene & Hensher, 2003) ja hierarkkinen Bayes-mallinnus (esim. Train 2001). Näillä menetelmillä on yleisesti tarkoitus ryhmitellä vastaajia aineistosta estimoitujen preferenssien mukaan tai laskea vastaajakohtaisia hyötyjä lainaamalla vastauksia muilta vastaajilta. Toinen mahdollisuus olisi lisätä malliin vastaajakohtaisia selittäjiä, kuten vaikkapa ikä tai sukupuoli. Vastaavilla tavoilla olisi myös mahdollista mallintaa ei mikään -vaihtoehdon valintaa. Kissanruoka-aineiston tapauksessa ne vastaajat, jotka eivät omista kissaa, valitsisivat joka kohdassa varmaankin ei mikään -vaihtoehdon.

Toinen vaihtoehto yksilökohtaisten preferenssien selvittämiseen olisi lisätä jokaisen vastaajan kuormaa lisäämällä valintatilanteita. Intuitiivisesti voi ajatella, että yhdeltä vastaajalta kysyttäessä tarpeeksi monta kertaa, vastaaja väsyy, ja valitsemisesta tulee satunnaisempaa. Käänteisesti voisi taas ajatella, että virhe minimoituisi, kun jokaista valintatilannetta kohden olisi vain yksi vastaaja. Tämä voi taas tulla kalliiksi, mikäli jokaiselle vastaajalle maksetaan kiinteä, koekohtainen palkkio tai jokaista vastaajaa kohden täytyy tehdä valmisteluja ennen koetta. Valintatilanteiden määrän ja vastaajien määrän suhteen on siis varmastikin olemassa jokin optimi, jonka eteen voisi olla hyvä tehdä tutkimusta.

Diskreetti valinta kokonaisuutena nojaa johdantokappaleen viitekehyksessä oleviin oletuksiin. Vaikka nämä oletukset ovatkin *normatiivisen*, rationaalisen toiminnan perusta, niin tuskinpa kovin moni nykyään edes väittää, että ne kelpaavat *deskriptiiviseksi* ihmiskuvaksi. Kaikissa tilanteissa näin ei ole edes keskimäärin (esim. Tversky & Simonson 1993). Näiden mekanismien ymmärtämistä vaaditaan, jotta ne voidaan tuoda osaksi malleja, joilla diskreettiä valintaa päätellään.

Lähteet

Agresti A. (2013) Categorical Data Analysis (3. painos), Wiley, New Jersey, 143.

Bishop, P. A., & Herron, R. L. (2015). Use and misuse of the likert item responses and other ordinal measures. *International journal of exercise science*, 8(3), 297.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, *78*(382), 316-331.

Haaijer, R., Kamakura, W., & Wedel, M. (2001). The'no-choice'alternative in conjoint choice experiments. *International Journal of Market Research*, 43(1), 93-106.

Hoffman, S. D., & Duncan, G. J. (1988). Multinomial and conditional logit discrete-choice models in demography. *Demography*, 25(3), 415-427.

Greene, W. H., & Hensher, D. A. (2003). A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological*, 37(8), 681-698.

Le, C. T., & Lindgren, B. L. (1988). Computational implementation of the conditional logistic regression model in the analysis of epidemiologic matched studies. *Computers and Biomedical Research*, 21(1), 48-52.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of mathematical psychology*, 15(3), 215-233.

Maddala, G. S. (1983), Limited-Dependent and Qualitative Variables in Economics, Cambridge University Press, New York, 60 – 61.

McLachlan G. J., Do K.-A. & Ambroise C (2004). Analyzing Microarray Gene Expression Data (1. painos), Wiley, New Jersey, 213.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153-157.

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems* (pp. 841-848).

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably

supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157-175.

Russell, S. ja Norvig, P. (2009), Artificial Intelligence: A Modern Approach (3. painos), Prentice Hall, New Jersey, 499, 505 ja 808.

Sy, J. P., & Taylor, J. M. G. (2001). Standard errors for the Cox proportional hazards cure model. *Mathematical and computer modelling*, 33(12-13), 1237-1251.

Train, K. (2001). A comparison of hierarchical Bayes and maximum simulated likelihood for mixed logit. University of California, Berkeley, 1-13.

Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management science*, 39(10), 1179-1189.

Liitteet

Liite 1 – Khii-toiseen yhteensopivuustesti ei mikään -vaihtoehtojen lukumäärille valintatilanteiden välillä

```
#Älypuhelin
> chisq.test(c(98,94,96,109,105,106,105,112,110))
    Chi-squared test for given probabilities

data:  c(14, 16, 18, 13, 21, 23, 21, 19, 27, 23, 22, 24)
X-squared = 9.7054, df = 11, p-value = 0.5571

#Kissanruoka
> chisq.test(c(14,16,18,13,21,23,21,19,27,23,22,24))
    Chi-squared test for given probabilities

data:  c(98, 94, 96, 109, 105, 106, 105, 112, 110)
X-squared = 3.185, df = 8, p-value = 0.9222
```

Liite 2 – Ehdollisten logit-mallien kertoimet eri ristiinvalidointiryhmien välillä.

	Väri					Näytön koko			Hinta (€)			Ei mikään
	Pun.	Vihreä	Sininen	Musta	Valkoinen	4″	5″	6″	299	199	99	
k=1	0	0,003 (0,119)	0,374 (0,114)	0,402 (0,113)	0,204 (0,117)	0	-0,89 (0,069)	-1,557 (0,089)	0	0,236 (0,08)	0,105 (0,082)	-0,341
k=2	0	0,005 (0,119)	0,36 (0,116)	0,452 (0,114)	0,125 (0,118)	0	-0,945 (0,07)	-1,595 (0,090)	0	0,265 (0,081)	0,129 (0,083)	-0,358
k=3	0	0,03 (0,119)	0,361 (0,115)	0,494 (0,113)	0,197 (0,117)	0	-0,897 (0,069)	-1,578 (0,089)	0	0,254 (0,08)	0,125 (0,082)	-0,346
k=4	0	-0,017 (0,119)	0,378 (0,115)	0,475 (0,114)	0,178 (0,117)	0	-0,924 (0,07)	-1,571 (0,089)	0	0,257 (0,08)	0,117 (0,082)	-0,352
k=5	0	0 (0,118)	0,343 (0,115)	0,450 (0,112)	0,150 (0,117)	0	-0,899 (0,069)	-1,563 (0,089)	0	0,24 (0,08)	0,116 (0,081)	-0,366
k=6	0	0,056 (0,118)	0,362 (0,115)	0,403 (0,114)	0,200 (0,117)	0	-0,922 (0,07)	-1,594 (0,090)	0	0,257 (0,081)	0,110 (0,082)	-0,351
k=7	0	0,047 (0,118)	0,37 (0,115)	0,449 (0,114)	0,196 (0,118)	0	-0,885 (0,069)	-1,594 (0,091)	0	0,227 (0,08)	0,114 (0,082)	-0,317
k=8	0	0,092 (0,119)	0,434 (0,116)	0,527 (0,113)	0,230 (0,119)	0	-0,906 (0,07)	-1,541 (0,089)	0	0,219 (0,08)	0,062 (0,082)	-0,298
k=9	0	0,069 (0,119)	0,41 (0,117)	0,478 (0,113)	0,226 (0,118)	0	-0,906 (0,07)	-1,597 (0,090)	0	0,294 (0,08)	0,126 (0,082)	-0,284
k=10	0	0,055 (0,12)	0,409 (0,117)	0,528 (0,114)	0,245 (0,119)	0	-0,939 (0,071)	-1,578 (0,090)	0	0,228 (0,082)	0,150 (0,082)	-0,247
Kaikki	0	0,034 (0,119)	0,38 (0,116)	0,466 (0,113)	0,195 (0,118)	0	-0,911 (0,07)	-1,577 (0,090)	0	0,248 (0,08)	0,115 (0,082)	-0,326

Älypuhelinaineisto, keskivirheet on esitetty sulkeissa

	Maku					Koostumus			Hinta (€)					Ei mikään
	Lohi	Tonnikala	Kana	Häränliha	Grillattu liha	hyytelö	kuiva	stick	8,99	5,30	4,29	3,90	2,50	
k=1	0	0,549 (0,111)	0,34 (0,114)	0,413 (0,113)	0,198 (0,116)	0	0,573 (0,075)	-0,688 (0,1)	0	-0,007 (0,105)	-0,112 (0,106)	-0,248 (0,109)	-0,184 (0,108)	-0,952
k=2	0	0,56 (0,111)	0,325 (0,114)	0,443 (0,112)	0,134 (0,118)	0	0,584 (0,075)	-0,638 (0,099)	0	0,083 (0,104)	-0,12 (0,108)	-0,207 (0,109)	-0,179 (0,109)	-0,934
k=3	0	0,561 (0,111)	0,358 (0,114)	0,433 (0,112)	0,189 (0,117)	0	0,575 (0,075)	-0,701 (0,1)	0	0,029 (0,104)	-0,12 (0,106)	-0,294 (0,11)	-0,272 (0,11)	-0,965
k=4	0	0,552 (0,109)	0,333 (0,112)	0,34 (0,112)	0,074 (0,117)	0	0,598 (0,075)	-0,668 (0,099)	0	-0,019 (0,104)	-0,154 (0,106)	-0,299 (0,109)	-0,227 (0,108)	-1,045
k=5	0	0,561 (0,11)	0,352 (0,114)	0,451 (0,113)	0,206 (0,116)	0	0,589 (0,075)	-0,6 (0,098)	0	-0,014 (0,104)	-0,116 (0,106)	-0,275 (0,109)	-0,194 (0,107)	-0,957
k=6	0	0,564 (0,11)	0,383 (0,113)	0,396 (0,113)	0,203 (0,117)	0	0,543 (0,074)	-0,687 (0,098)	0	-0,017 (0,104)	-0,141 (0,106)	-0,259 (0,109)	-0,15 (0,106)	-1,019
k=7	0	0,588 (0,11)	0,351 (0,113)	0,389 (0,113)	0,221 (0,116)	0	0,558 (0,075)	-0,696 (0,099)	0	-0,009 (0,104)	-0,145 (0,106)	-0,263 (0,109)	-0,218 (0,108)	-0,997
k=8	0	0,598 (0,111)	0,381 (0,114)	0,443 (0,113)	0,215 (0,117)	0	0,614 (0,075)	-0,625 (0,099)	0	-0,005 (0,103)	-0,143 (0,105)	-0,319 (0,109)	-0,236 (0,108)	-0,97
k=9	0	0,559 (0,11)	0,349 (0,114)	0,402 (0,113)	0,174 (0,117)	0	0,535 (0,075)	-0,67 (0,098)	0	-0,004 (0,104)	-0,147 (0,106)	-0,275 (0,11)	-0,196 (0,108)	-0,997
k=10	0	0,536 (0,11)	0,304 (0,113)	0,395 (0,112)	0,134 (0,117)	0	0,574 (0,075)	-0,65 (0,099)	0	0,009 (0,104)	-0,138 (0,107)	-0,294 (0,11)	-0,161 (0,107)	-0,977
Kaikki	0	0,563 (0,11)	0,348 (0,113)	0,41 (0,113)	0,175 (0,117)	0	0,574 (0,075)	-0,662 (0,099)	0	0,005 (0,104)	-0,134 (0,106)	-0,273 (0,109)	-0,202 (0,108)	-0,981

Kissanruoka-aineisto, keskivirheet on esitetty sulkeissa

Liite 3 – Visual Basic -syntaksi ei mikään -vaihtoehdon hyötyjen simuloimiseksi

```
1.    Sub eiMikaanHyotyjenSimulointi()
2.
3.    Dim i As Long
4.    a = 0
5.    b = 0
6.    c = 0
7.    d = 0
8.    e = 0
9.    f = 0
10.   g = 0
11.   h = 0
12.   j = 0
13.   k = 0
14.
15.   Do Until i > 1000000
16.   a = a + ActiveCell.Offset(-4, 0)
17.   b = b + ActiveCell.Offset(10, 0)
18.   c = c + ActiveCell.Offset(24, 0)
19.   d = d + ActiveCell.Offset(38, 0)
20.   e = e + ActiveCell.Offset(52, 0)
21.   f = f + ActiveCell.Offset(66, 0)
22.   g = g + ActiveCell.Offset(80, 0)
23.   h = h + ActiveCell.Offset(94, 0)
24.   j = j + ActiveCell.Offset(108, 0)
25.   k = k + ActiveCell.Offset(122, 0)
26.
27.   ActiveSheet.EnableCalculation = False
28.   ActiveSheet.EnableCalculation = True
29.   i = i + 1
30.
31.   Loop
32.   ActiveCell.Value = a / i
33.   ActiveCell.Offset(14, 0).Value = b / i
34.   ActiveCell.Offset(28, 0).Value = c / i
35.   ActiveCell.Offset(42, 0).Value = d / i
36.   ActiveCell.Offset(56, 0).Value = e / i
37.   ActiveCell.Offset(70, 0).Value = f / i
38.   ActiveCell.Offset(84, 0).Value = g / i
39.   ActiveCell.Offset(98, 0).Value = h / i
40.   ActiveCell.Offset(112, 0).Value = j / i
41.   ActiveCell.Offset(126, 0).Value = k / i
42.
43.   End Sub
```

Liite 4 - Visual Basic -koodi Kissanruoka-aineiston valitsimen kontrolloimiseksi, ja todennäköisyyksien laskemiseksi

```
1.    Sub kissanruokaNBC()
2.    'Aloitetaan
3.    10
4.
5.    'Katsotaan ollaanko jo lopussa
6.    If ActiveCell.Offset(0, -1).Value <> "" Then
7.
8.
9.    'Asetetaan parametrit
10.   Range("Paneeli!D2") = ActiveCell.Offset(0, -1).Value
11.
12.   Range("Paneeli!B6") = ActiveCell.Offset(0, -13).Value
13.   Range("Paneeli!B7") = ActiveCell.Offset(0, -12).Value
14.   Range("Paneeli!B8") = ActiveCell.Offset(0, -11).Value
15.
16.   Range("Paneeli!E6") = ActiveCell.Offset(0, -10).Value
17.   Range("Paneeli!E7") = ActiveCell.Offset(0, -9).Value
18.   Range("Paneeli!E8") = ActiveCell.Offset(0, -8).Value
19.
20.   Range("Paneeli!H6") = ActiveCell.Offset(0, -7).Value
21.   Range("Paneeli!H7") = ActiveCell.Offset(0, -6).Value
22.   Range("Paneeli!H8") = ActiveCell.Offset(0, -5).Value
23.
24.   Range("Paneeli!K6") = ActiveCell.Offset(0, -4).Value
25.   Range("Paneeli!K7") = ActiveCell.Offset(0, -3).Value
26.   Range("Paneeli!K8") = ActiveCell.Offset(0, -2).Value
27.
28.   'Luetaan lopputulos
29.   ActiveCell.Offset(0, 0).Value = Range("Paneeli!B9")
30.   ActiveCell.Offset(0, 1).Value = Range("Paneeli!E9")
31.   ActiveCell.Offset(0, 2).Value = Range("Paneeli!H9")
32.   ActiveCell.Offset(0, 3).Value = Range("Paneeli!K9")
33.   ActiveCell.Offset(0, 4).Value = Range("Paneeli!K12")
34.
35.   'Siirrytään alempaan soluun
36.   ActiveCell.Offset(1, 0).Activate
37.
38.   GoTo 10
39.   Else: End If
40.
41.   End Sub
```

Liite 5 - Visual Basic -koodi Älypuhelin-aineiston valitsimen kontrolloimiseksi, ja todennäköisyyksien laskemiseksi

```
1.    Sub alypuhelinNBC()
2.    'Aloitetaan
3.    10
4.
5.    'Katsotaan ollaanko jo lopussa
6.    If ActiveCell.Offset(0, -1).Value <> "" Then
7.
8.
9.    'Asetetaan parametrit
10.   Range("Paneeli!D2") = ActiveCell.Offset(0, -1).Value
11.
12.   Range("Paneeli!B6") = ActiveCell.Offset(0, -12).Value
13.   Range("Paneeli!B7") = ActiveCell.Offset(0, -11).Value
14.   Range("Paneeli!B8") = ActiveCell.Offset(0, -10).Value
15.
16.   Range("Paneeli!E6") = ActiveCell.Offset(0, -9).Value
17.   Range("Paneeli!E7") = ActiveCell.Offset(0, -8).Value
18.   Range("Paneeli!E8") = ActiveCell.Offset(0, -7).Value
19.
20.   Range("Paneeli!H6") = ActiveCell.Offset(0, -6).Value
21.   Range("Paneeli!H7") = ActiveCell.Offset(0, -5).Value
22.   Range("Paneeli!H8") = ActiveCell.Offset(0, -4).Value
23.
24.   'Luetaan lopputulos
25.   ActiveCell.Offset(0, 0).Value = Range("Paneeli!B18")
26.   ActiveCell.Offset(0, 1).Value = Range("Paneeli!B19")
27.   ActiveCell.Offset(0, 2).Value = Range("Paneeli!B20")
28.   ActiveCell.Offset(0, 3).Value = Range("Paneeli!B21")
29.
30.   'Siirrytään alempaan soluun
31.   ActiveCell.Offset(1, 0).Activate
32.
33.   GoTo 10
34.   Else: End If
35.
36.   End Sub
```

Liite 6 – McNemarin testi ja binomitesti mallien välillä R-ohjelmalla

```
> #McNemarin testi
> mcnemar.test(matrix(c(1281,12,8,2305),2,2), correct = FALSE)

        McNemar's Chi-squared test

data:  matrix(c(1281, 12, 8, 2305), 2, 2)
McNemar's chi-squared = 0.8, df = 1, p-value = 0.3711

> #Binomitesti
> binom.test(8, 20)

        Exact binomial test

data:  8 and 20
number of successes = 8, number of trials = 20, p-value = 0.5034
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1911901 0.6394574
sample estimates:
probability of success
                   0.4
```

Liite 7 – McNemarin testi ja binomitesti mallien välillä R-ohjelmalla, kun sellaiset vastaajat on poistettu, jotka valitsivat ei mikään -vaihtoehdon joka kerralla

```
> #McNemarin testi
> mcnemar.test(matrix(c(1281,12,7,1688),2,2), correct = FALSE)

        McNemar's Chi-squared test

data:  matrix(c(1281, 12, 7, 1688), 2, 2)
McNemar's chi-squared = 1.3158, df = 1, p-value = 0.2513

> #Binomitesti
> binom.test(7, 19)

        Exact binomial test

data:  7 and 19
number of successes = 7, number of trials = 19, p-value = 0.3593
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1628859 0.6164221
sample estimates:
probability of success
             0.3684211
```